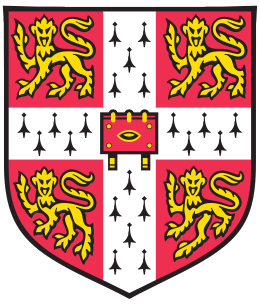
Methods for Data Management in Multi-Centre MRI Studies and Applications to Traumatic Brain Injury



Stefan Oliver Winzeck

Darwin College

Department of Medicine

University of Cambridge

Supervisor: Prof. David Menon

Advisor: Dr. Marta Correia

This dissertation is submitted for the degree of

Doctor of Philosophy

November 2020

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation contains less than 60,000 words excluding appendices, bibliography, footnotes, figures, tables and equations.

Abstract

PhD Thesis - Stefan Winzeck

Methods for Data Management in Multi-Centre MRI Studies and Applications to Traumatic Brain Injury

Neuroimaging studies are becoming increasingly bigger, and multi-centre collaborations to collect data under similar protocols, but different scanning sites, are now commonplace. However, with increasing sample size the complexity of databases and the entailed data management as well as computational burden are growing. This thesis aims to highlight and address challenges faced by large multi-centre *magnetic resonance imaging* (MRI) studies. The methods implemented are then applied to *traumatic brain injury* (TBI) data. Firstly, a pre-processing pipeline for both anatomical and diffusion MRI was proposed, that allows for a high throughput of MRI scans. After describing the choices for processing tools, the performance of the integrated quality assurance was assessed based on the results from a large multi-centre dataset for TBI. Secondly, the applicability of the pipelines for processing *mild TBI* (mTBI) data from three sites was shown in a case study. For this, volumetric and diffusion metrics in the acute phase are analysed for their prognostic potential. Furthermore, the cohort was examined for longitudinal changes. Thirdly, independent scan-rescan datasets are examined to gain a better understanding of the degree of reproducibility which can be achieved in imaging studies. This involves analysing the robustness of brain parcellations based on structural or diffusion imaging. The effect of using different MRI scanners or imaging protocols was also assessed and discussed. Fourthly, sources of diffusion MRI variability and different approaches to cope with these are reviewed. Using this foundation, state-of-the art methods for diffusion MRI harmonisation were compared against each other using both a benchmark dataset and mTBI cohort. Lastly, a solution to localise brain lesions was proposed. Its implications for lesion analysis, are assessed in the light of an application to a more severe TBI patient cohort, imaged on two different scanners. Furthermore, a lesion matching algorithm was introduced to automatically examine lesion evolution with time post-injury. In summary, this thesis explored different options for MRI data analysis in the context of large multi-centre studies. Different approaches are studied and compared using a number of different MRI datasets, including scan-rescan data across different MRI scanners and imaging protocols. The potential of the optimised solutions was illustrated through applications to TBI data.

Acknowledgements

First of all, I would like to thank Prof. David Menon for giving me the opportunity to pursue my PhD in his lab, providing me with the needed resources and sharing with me his professorial point of view in many insightful discussions. I also thank Dr. Virginia Newcombe for her engagement and her valuable input regarding TBI.

Most of all, I thank Dr. Marta Correia for guiding me with kindness and competence over the past years. Teaching me critical thinking and being a role model for me, I truly think this thesis would not have been possible without her.

Furthermore, I thank Dr. Emmanuel Stamatakis for his advice and collaborative work. Many thanks also to Dr. Evgenios Kornaropoulos, as we formed an engineering team among the clinicians. Many thanks to Kevin Kunzmann for his general advice on statistical analysis. I am also grateful to Dr. Ellen Carroll, Dot Chatfield, Liana Dordai and Faye Forsyth as well as Anne Manktelow and Joanne Outtrim for their efforts for patient recruitment, data acquisition and management. Since the HPHI was indispensable for this thesis, I would like to thank Dr. Guy Williams and Paul Browne for their support. I also thank Jane Miller - for taking care of all the non-research related issues and for looking out for me - and my fellow engineer Abhishek Dixit.

Although not directly linked to this thesis, I thank Dr. Ona Wu for her support and the opportunities provided throughout the years, which made me a better researcher. Likewise, I would like to thank Prof. Mauricio Reyes for the fruitful and engaging collaboration.

I believe, next to every ambitious man stands at least one strong woman, which is why I am thankful to my friends Eirini, Eugenia, Jemma and Lotti for being my safety net. Lastly, I thank my family for the unconditional support.

Contents

1 Introduction **1**

1.1 Neuroimaging: A Field Moves Towards *Big Data* 1

1.2 Design of Big Neuroimaging Studies 2

1.2.1 Large Single-Centre Studies 2

1.2.2 Retrospective Multi-Centre Studies 3

1.2.3 Prospective Multi-Centre Studies 4

1.3 Challenges 5

1.3.1 Data Management 5

1.3.2 Confounding Factors 5

1.3.3 Data Harmonisation 6

1.4 Overview of Thesis 7

2 Data Analysis Pipelines **9**

2.1 Requirements and Concept of Processing Pipelines 9

2.2 Data Acquisition 10

2.3 Pipeline for Structural Magnetic Resonance Images 12

2.3.1 Processing Modules 14

2.3.2 Quality Control Metrics 21

2.4 Pipeline for Diffusion Magnetic Resonance Images 26

2.4.1 Processing Modules 27

2.4.2 Quality Control Metrics 34

2.5 Discussion 39

2.5.1 Database Management and Quality Control 39

2.5.2 The Pipeline - An Ever Evolving Process 41

2.5.3 Future Developments 41

2.5.4 Integrated Lesion Segmentation 43

2.6 Chapter Summary 44

3 Application to Mild TBI **45**

3.1 Introduction . . . 45

3.1.1 Brief Introduction to Traumatic Brain Injury . . . 45

3.1.2 Role of Neuroimaging for Mild Traumatic Brain Injury . . . 47

3.1.3 Related Work for Analysis of Mild TBI MRI Data . . . 48

3.1.4 Aims . . . 52

3.2 Data Acquisition, Processing & Curation . . . 53

3.2.1 Databases . . . 53

3.2.2 Specifications of MRI Processing . . . 54

3.2.3 Data Curation Prior to Analysis . . . 55

3.3 Experiment Setup . . . 56

3.3.1 Data Categorisation . . . 56

3.3.2 Region Selection for Analysis . . . 57

3.3.3 General Statistical Analysis . . . 58

3.3.4 Site-Specific Biases . . . 59

3.3.5 Acute MRI Differences between Controls and Patient Groups . . . 59

3.3.6 Prognostic Value of Acute MRI for Mild TBI Outcome . . . 60

3.3.7 Longitudinal Analysis . . . 61

3.4 Results . . . 63

3.4.1 Site-Specific Biases . . . 63

3.4.2 Acute Differences between Controls and Patients . . . 64

3.4.3 Prognostic Value of Acute MRI . . . 69

3.4.4 Longitudinal Analysis . . . 71

3.5 Discussion . . . 72

3.5.1 General Challenges in Database Comparability . . . 72

3.5.2 Biases Across Imaging Sites . . . 73

3.5.3 Differences between Patients in Comparison to Controls . . . 74

3.5.4 Prognosis Based on Acute MRI . . . 75

3.5.5 Longitudinal Findings in Mild TBI Patients . . . 76

3.5.6 Heterogeneity of the TBI Cohort . . . 78

3.6 Chapter Summary . . . 78

4 Reproducibility of MRI Metrics **80**

4.1 Introduction . . . 80

4.1.1 Variability of Anatomical Brain Segmentation . . . 80

4.1.2 Reproducibility of Diffusion Magnetic Resonance Imaging . . . 81

4.1.3 Overview & Aims . . . 86
4.2 Data & Methods . . . 87
4.2.1 Databases . . . 87
4.2.2 Experiment Setup . . . 88
4.2.3 Evaluation Metrics . . . 92
4.3 Results . . . 93
4.3.1 Reproducibility of Anatomical Brain Parcellation . . . 93
4.3.2 Variation of White Matter Region Segmentation . . . 96
4.3.3 Inter-Scanner Differences of DTI Metrics . . . 99
4.3.4 Impact of Acquisition Protocol on Fibre Tract Segmentation . . . 102
4.3.5 Acquisition Protocol Specific Differences in DTI Metrics . . . 102
4.3.6 Comparability of Multi-Centre Data Acquisition . . . 106
4.3.7 Variation of DTI Metrics for Multi-Centre Data . . . 107
4.4 Discussion . . . 114
4.4.1 Parcellation of Structural Brain Scans . . . 114
4.4.2 Segmentation of White Matter Tracts . . . 115
4.4.3 Variability of DTI Metrics . . . 116
4.4.4 The Difficulty of Measuring Variability . . . 118
4.5 Chapter Summary . . . 119

5 Harmonisation of DWI for Multi-Centre Studies 120
5.1 Introduction . . . 120
5.1.1 Sources of Variation in Diffusion MRI . . . 120
5.1.2 Spherical Harmonics & Rotation Invariant Features . . . 121
5.1.3 Related Work . . . 121
5.1.4 Advantages and Limitations of Existing Methods . . . 127
5.1.5 Aims . . . 128
5.2 Data & Methods . . . 129
5.2.1 Databases . . . 129
5.2.2 Benchmarking Implementations of Harmonisation. . . . 129
5.2.3 Inter-Scanner Variation for CENTER-TBI Substudy . . . 133
5.2.4 Denosing of RISH Feature Scaling Maps . . . 133
5.2.5 Subject Selection for RISH Feature Scaling Maps . . . 134
5.2.6 Evaluation of Harmonisation of CENTER-TBI SH Images . . . 135
5.2.7 Data Harmonisation of CENTER-TBI DTI Metrics . . . 136
5.2.8 Impact of Data Harmonisation on Mild TBI . . . 136

5.3 Results . . . 137
5.3.1 Comparison of Selected Harmonisation Methods . . . 137
5.3.2 Variation in CENTER-TBI Substudy . . . 138
5.3.3 Impact of Denoising on RISH Feature Scaling Maps . . . 138
5.3.4 Scan Selection and Weighting for Scaling Maps . . . 144
5.3.5 Harmonisation of CENTER-TBI SH Images . . . 147
5.3.6 Evaluation of Harmonisation for CENTER-TBI DTI Metrics . . . 148
5.3.7 Impact of Data Harmonisation on Mild TBI . . . 152
5.4 Discussion . . . 153
5.4.1 Potential and Limitations of Harmonisation Methods . . . 153
5.4.2 Enhancement of Scaling Maps Through Post-Processing . . . 155
5.4.3 Application to Traumatic Brain Injury Data . . . 156
5.4.4 Future Work . . . 157
5.5 Chapter Summary . . . 158

6 Lesions Analysis in Severe TBI 159

6.1 Introduction . . . 159
6.1.1 Motivation from a Clinical Research Perspective . . . 160
6.1.2 General Concept for Lesion Localisation . . . 161
6.1.3 Assessment of Lesion Progression . . . 163
6.1.4 Aims . . . 163
6.2 Data & Methods . . . 164
6.2.1 Severe TBI Database . . . 164
6.2.2 Localisation of Lesions . . . 165
6.2.3 Longitudinal Lesion Matching . . . 166
6.3 Results . . . 167
6.3.1 Group-Wise Lesion Burden Across Scanners . . . 167
6.3.2 Lesion Volume Progression After TBI . . . 169
6.3.3 Subject-Wise Cross-Scanner Comparison of Lesions . . . 172
6.3.4 Subject-Wise Lesion Characteristics on Longitudinal Scans . . . 173
6.3.5 Automated Lesion Matching . . . 176
6.4 Discussion . . . 182
6.4.1 Summary of Findings . . . 182
6.4.2 Limitations of Study . . . 185
6.4.3 Future Work . . . 186
6.5 Chapter Summary . . . 188

7 Summary **189**

7.1 Summary of Findings 189

7.2 Limitations & Future Directions 192

7.3 Conclusion 195

Appendix A **196**

Nomenclature

l_{max}	Highest order of spherical harmonics
3D	Three dimensional
3T	Three Tesla
ABIDE	Autism brain imaging data exchange
AMB	Anterior mid-body
ANOVA	Analysis of variance
AP	Anisotropic power map
ATR	Anterior thalamic radiation
BBB	blood–brain barrier
BET	Brain extraction tool (from FSL)
Cam-CAN	Cambridge centre for ageing and neuroscience
CC	Corpus callosum
CENTER-TBI	Collaborative European neuro-trauma effectiveness research in traumatic brain injury
CG	Cingulum
CNN	Convolutional neural network
CNR	Contrast-to-nose ratio
ComBat	Combined association test
CSF	Cerebrospinal fluid
CST	Corticospinal tract

CSV	Comma-separated values
CT	Computer tomography
CV	Coefficients of variance
DiReCT	Diffeomorphic registration-based cortical thickness
DKI	Diffusion kurtosis imaging
DPI	Days post-injury
DTI	Diffusion tensor imaging
DWI	Diffusion weighted imaging/images
ENIGMA	Enhancing neuroimaging genetics through meta-analysis
EPI	Echo-planar imaging
FA	Fractional anisotropy
FCSNet	Fully-convolutional shuffling network
FDR	False discovery rate
FLAIR	Fluid attenuated inversion recovery
FOV	Field of view
FW	Free water
GE	General Electrics
GLM	Generalised linear model
GM	Grey matter
GOSE	Extended Glasgow Outcome Scale
GOS	Glasgow outcome scale
GRE	Gradient echo
HCP	Human connectome project
IDP	Image derived phenotype
IFO	Inferior fronto-occipital tract

ILF	Inferior longitudinal fascicle
IQR	Interquartile range
JHU	John Hopkins University
MALP-EM	Multi-atlas label propagation with expectation maximisation based refinement
MD	Mean diffusivity
MI	Mutual information
MK	Mean kurtosis
MNI	Montreal Neurological Institute
MoM	Method of moments
MPPCA	Marcenko-Pastur Principal component analysis
MPRAGE	Magnetisation-prepared rapid acquisition with gradient echo
MRI	Magnetic resonance imaging
MR	Magnetic resonance
MSE	Mean square error
mTBI	Mild traumatic brain injury
MUSHAC	Multi-shell diffusion MRI harmonisation and enhancement challenge
NCC	Normalised cross correlation
NMI	Normalised mutual information
OASIS	Open access series of im studies
ODF	Orientation distribution function
PCA	Principal component analysis
PD	Proton density
PIS	Physically implausible signal
PMB	Posterior mid-body
PNC	Philadelphia neurodevelopmental cohort

QC	Quality control
ResBlock	Residual block
RF	Randomised forest
RISH	Rotation invariant spherical harmonics
rm-ANOVA	Analysis of variance for repeated measurements
RMSE	Root mean square error
ROI	Region of interest
SHNet	Spherical harmonic network
SHResNet	Spherical harmonic residual network
SH	Spherical harmonics
SLF	Superior longitudinal fascicle
SNR	Signal-to-nose ratio
std	Standard deviation
SWI	Susceptibility weighted im
T1w	T1-weighted
T2w	T2-weighted
TBI	Traumatic brain injury
TBSS	Tract-based spatial statistics
TE	Echo time
TI	Inversion time
TRACK-TBI	Transforming Research and clinical knowledge in traumatic brain injury
TR	Repetition time
UF	Uncinate fascicle
VAE	Variational auto-encoder
VBM	Voxel-based morphometry
WM	White matter

List of Figures

2.1 Schematic Overview of Pipeline for Structural MRI 14
2.2 Example of Neck Cropping 15
2.3 Example Comparison of Brain Masking Algorithms 16
2.4 Example of T1w Image and Computed Feature Maps from a Healthy Subject 17
2.5 Quality Control of Brain Extraction for CENTER-TBI Database 22
2.6 Quality Control of Spatial Normalisation for CENTER-TBI Database 23
2.7 Quality Control of Coregistration of CENTER-TBI Database 26
2.8 Overview of Pipeline for Diffusion MRI 27
2.9 Example of Multi-Shell DWI Denoising from Cam-CAN Database 29
2.10 Overview of Diffusion Parameter Maps for an Individual Subject 32
2.11 Distribution of SNR across Centres 35
2.12 Quality Controls for DWI Brain Masking 37
2.13 Quality Control for DWI Head Motion for CENTER-TBI Subjects 38
2.14 Assessment of DTI Coregistration 39

3.1 Overview of Age and Scan Time Distribution 57
3.2 Examples of Regional Volume Differences Across Sites 63
3.3 Regional FA Differences in Control Subjects 65
3.4 Examples of Regional FA Differences 67

4.1 Distribution of Volumes from MALP-EM Regions Deviating Across Scanners 95
4.2 Distribution of Volumes from TractSeg Regions Deviating Across Centre . . . 98
4.3 Differences of Average FA Intensities within Selected TractSeg ROIs 100
4.4 Differences of Average MD Intensities within Selected TractSeg ROIs 101
4.5 Reputability of Tract Volumes for Different DWI Acquisitions 103
4.6 Relationship Between CV of DTI Metrics and Volumes of TractSeg Parcellation 104
4.7 Distribution of Regional Coefficients of Variances Across Different Protocols . 105
4.8 Distribution of Quality Control Metrics for Control Subjects 108

4.9 Intra- and Inter-Scanner Distribution of CV within ROIs 111
4.10 Coefficients of Variation Maps for FA and MD 114

5.1 Schematic Architecture of Convolutional Neural Networks 132
5.2 Comparison of Global RISH Feature Differences after Harmonisation 137
5.3 Comparison of Regional RISH Feature Differences after Harmonisation 140
5.4 Intra- and Inter-Scanner Variation for CENTER-TBI Cambridge Subset . . . 141
5.5 Effect of Denoising of Scaling Maps for Data Harmonisation 143
5.6 Example of Scaling Maps of CENTER-TBI Subject 145
5.7 Impact of Selection of Subjects on Image Harmonisation 147
5.8 Comparison of Harmonisation Methods on CENTER-TBI Controls 148
5.9 Comparison of DTI Metrics Before and After Harmonisation 150
5.10 Comparison of Regional DTI Metrics Before and After Harmonisation 150
5.11 Comparison of Means DTI Metrics in Controls Before and After Harmonisation 152
5.12 Regional Mean FA and MD Before and After Harmonisation for mTBI Patients 153

6.1 Problematic of Automated Brain Parcellation in Presence of Lesions 162
6.2 Comparison of Original and Projected Parcellation 165
6.3 Schematic Overview of Lesion Matching Between Longitudinal Scans 168
6.4 Distribution of FLAIR Contusion Core and Oedema 169
6.5 Progression of Contusion Core and Oedema after TBI 171
6.6 Lesion Overlap at Different Phases Post-Injury 174
6.7 Longitudinal Comparison of Contusion Lesion Volumes Post-Injury 175
6.8 Examples of Matched Contusions between Hyper-Acute and Acute Scans . . 178

A.1 TractSeg Volume Distribution for CENTER-TBI 199

List of Tables

2.1 Example of Acquisition Parameters of CENTER-TBI Database 13
2.2 Acquisition Parameters of Scan-Rescan Database 13
2.3 NCC for Coregistered Sequences 25
2.4 Quality Metrics for Head Motion During DTI Acquisition 37

3.1 Number of Subjects and Scans of the mTBI Databases 56
3.2 Overview of Available Data for MRI Analysis 61
3.3 Overview of Available Data for Longitudinal Analysis 62
3.4 Overview of ROIs with Different FA in Acute Phase Across Sites 66
3.5 Overview of ROIs with Different Diffusion Metric in Acute Phase 68

4.1 Overview of Included Data of CENTER-TBI Controls 92
4.2 MALP-EM Volumes for Tissue Compartments 94
4.3 P-Values After FDR Correction for Comparison of MALP-EM ROI Volumes 96
4.4 P-Values After FDR Correction for Comparison of TractSeg ROI Volumes . . 97
4.5 Average CV of DTI Metrics for JHU and TractSeg ROIs 106
4.6 Quality Controls Metrics Across Scanner for CENTER-TBI Controls 109
4.7 Average of CV Maps within JHU ROIs 113

5.1 Overview of Cambridge Data for CENTER-TBI 130
5.2 Global RMSE for Different Harmonisation Methods 139
5.3 Global and Regional Variation within and Across Scanners 142
5.4 Impact of Denoising of Scaling Maps on Harmonisation between Scanners . . 144
5.5 Impact of Subject Selection and Weighting on Harmonisation 146
5.6 Comparison of Harmonisation Methods on Control Subjects 149
5.7 Global and Regional RMSE for FA and MD Within and Across Scanners . . 151

6.1 Overview of Lesion Volumes for Different Time Windows after TBI 170
6.2 Lesion Characteristics After TBI 176

6.3 Volume Changes of Matched TBI Contusion Clusters 179
6.4 Patient-Wise Changes of Contusion Core and Oedema Volumes 181

A.1 P-Values After FDR Correction for Comparison of TractSeg ROI Mean FA . 196
A.2 P-Values After FDR Correction for Comparison of TractSeg ROI Mean MD . 197
A.3 P-Values After FDR Correction for Comparison of JHU ROI Mean DTI Metrics 198
7.4 Effect Size of Harmonisation Methods on Control Subjects from CENTER-TBI. Cohen's effect size d was computed between global RMSE values. Effect sizes can be small ($d = \pm 0.2$), medium ($d = \pm 0.5$) or large ($d = \pm 0.8$). 199

Chapter 1

Introduction

1.1 Neuroimaging: A Field Moves Towards *Big Data*

When the first *magnetic resonance* (MR) images of a live human subject were acquired in the late 1970's, new possibilities to study human anatomy and physiology were opened up. Shortly after, with commercially available MRI scanners, the field of neuroimaging was born [59]. Most studies in the first three decades have been limited to small number of subjects, but over the years medical imaging has seen many new advances in hardware and pulse sequence design as well as image reconstruction algorithms. With MRI scanners becoming cheaper, faster [248] and more accessible [76] over time, and with the new emerging fields - such as the ultra-high field MRI [201] - more scans are collected than ever. This trend is further enforced as studies nowadays usually acquire multi-modal images to examine anatomy, function and connectivity of the healthy and diseased brain. Besides the number of obtained scans, their capacity has progressively transformed to capture more detailed information about the brain. For example, the earliest functional image time series sampled total volumes at a four second rate, but with better imaging technologies this interval was more than halved while slice resolution increased [252]. Similarly, for *diffusion tensor imaging* (DTI) for which initially only six volumes, each sensitised along a different gradient direction, were acquired [16]. To better examine *white matter* (WM) fibre orientation, the angular resolution was refined by increasing the number of diffusion directions [50, 279]. For example, diffusion spectral imaging has been introduced to resolve more than 500 diffusion directions while keeping common volume size and spatial resolution [59, 252]. Although useful to quantify the micro-structural integrity of WM, diffusion spectral imaging is limited to research applications and today's clinical studies often include around 60 diffusion volumes. In addition to the general increasing size of individual datasets through multi-

modal image acquisition, there is a trend to collect larger numbers of subjects to improve statistical power. Large-scale databases enable the detection of subtle effects that would not be statistically observable in smaller groups [233]. Acknowledging this need for larger datasets, both population and multi-centre studies have emerged throughout the past few years. While smaller studies remain important to look at very specific cohorts in future, the shift towards larger databases in the field of neuroimaging is clearly observable [272]. Such *big* databases can be shaped by either a large number of subjects (i.e. >1,000) scanned or the extensive data acquisition with advanced scanning protocols (e.g. *Human Connectome Project*[1] [HCP] *Adult Diffusion* sequence with several shells and high angular resolution amasses more than three gigabytes per scan) [233]. Both have the potential to improve image analysis and foster more generalisable research. Eventually, larger neuroimaging studies will help to understand specific patient cohorts as well as disease development in a general healthy populations, however, they also come with many challenges for data management, processing and analysis.

1.2 Design of Big Neuroimaging Studies

Collecting big data for neuroimaging studies can be difficult, which is why different approaches have been established to increase the size of databases. They can be broadly categorised as single-centre studies and retrospective as well as prospective multi-centre studies. All of them come with different advantages and drawbacks.

1.2.1 Large Single-Centre Studies

In a best case scenario, subject recruitment and data collection is centralised at one imaging site. With one MR protocol used ideally on one scanner consistently, the acquisition related variability can be greatly reduced. Besides this, the administrative effort and costs can be kept to a minimum as no coordination between sites is required. One example for a large-scale database is the single-site *Cambridge Centre for Ageing and Neuroscience*[2] (Cam-CAN) study that collected neuroimaging and cognitive data for several hundred of healthy subjects across a wide age range to research healthy ageing [226]. Another example is the HCP which accumulated multi-parametric neuroimaging data on one scanner from 1,200 healthy volunteers over the span of five years [251]. A little more specific is the *Philadelphia Neurodevelopmental Cohort*[3] (PNC) initiative. For this more than 1,400

[1] `www.humanconnectomeproject.org`
[2] `www.cam-can.org`
[3] `www.med.upenn.edu/bbl/philadelphianeurodevelopmentalcohort.html`

adolescents underwent multi-modal MRI to examine cognitive development and its link to psychiatric illness. To form a big database, inclusion criteria were kept to a minimum and no screening for specific psychiatric or other medical disorders took place at the recruitment stage [223]. All three studies have in common, that they mainly focus on healthy volunteers or have relative wide inclusion criteria for subject recruitment.

1.2.2 Retrospective Multi-Centre Studies

Although acquiring imaging data at a single site poses much less challenges for data organisation and harmonisation, one big shortcoming may be the availability of patients. Multi-site studies have the potential to amass a large dataset from a specific patient cohort, which otherwise would not be obtainable at a single site. Besides increasing the sample size, combined datasets from multiple centres, various cities and different countries bear the potential to capture the full disease profile. Since many clinical facilities collect data from specific patient cohorts for diagnosis and treatment, a great opportunity to increase sample size is to aggregate such legacy datasets. For a small number of partnering sites, this usually requires little organisational effort. However, such collaborations can be limited to datasets that were acquired with similar imaging techniques to allow any combined analysis [249]. Projects could range from two research groups, that agree to share data, up to large-scale collaborations incorporating multiple sites. A special case are consortia which aim to aggregate data from many sites to provide a large database of certain cohorts. For instance, the *Autism Brain Imaging Data Exchange*[4] (ABIDE) project which gathered and openly provides resting-state functional MRI for more than 1,000 subjects (fairly even split between autistic individuals and age-matched controls) from over 24 international brain imaging sites [54]. Another prominent example is the *Enhancing Neuroimaging Genetics through Meta-Analysis*[5] (ENIGMA) consortium - a growing number of scientists across 340 institutions in 35 countries - investigating the genetic impact on brain structure. Pooling imaging and genetics data from global existing studies amplifies the sample sizes by many folds. This fosters research of brain abnormalities in more than 20 major cognitive, psychiatric and neurogenetic disorders (e.g. depression or schizophrenia) [18, 242]. Since imaging data are acquired independently and combined retrospectively in a meta-analysis, the efforts for prior coordination are low, however, with the caveat of a highly heterogeneous database.

[4] `www.fcon_1000.projects.nitrc.org/indi/abide`
[5] `www.enigma.usc.edu`

1.2.3 Prospective Multi-Centre Studies

Multi-centre imaging studies require the most planning and substantial effort to coordinate the data collection. This is reaching from equalising imaging equipment and methods across sites to building a sufficient infrastructure to centralise and share data. Such studies are very expensive due to planning overhead, however, with the benefit of minimising the variation in inclusion criteria, scanner calibration and imaging protocols, which facilitates the pooling of data for a combined analysis [82, 249]. Such large databases are particularly important if the cohort under investigation exhibits a vast heterogeneity, such as for example TBI patients, who suffer from injuries of different severities and experience a wide spectrum of symptoms. A well established prospective multi-centre project to investigate TBI on a large scale is the *Transforming Research and Clinical Knowledge in Traumatic Brain Injury*[6] (TRACK-TBI) study. It aimed to gather neuroimaging data, including MRI two weeks after injury, at four sites in the USA. All centres implemented equivalent imaging protocols on MRI scanners from three different vendors [274]. Similarly, the more extended *Collaborative European Neuro-Trauma Effectiveness Research in Traumatic Brain Injury*[7] (CENTER-TBI) study has been set up as a prospective collaboration between 22 countries (ca. 80 sites) across Europe and Israel with the aim to collectively gather extensive data to improve characterisation of TBI. This included the collection of longitudinal clinical and epidemiological information as well as blood biomarkers. Some of the sites have been contributing to an extended data acquisition, which included multi-contrast MRI data collected up to two years post-injury [162]. In order to minimise acquisition related differences, site-specific MR protocols were designed with similar parameters. To ensure the image quality and provide a baseline, scanning protocols were validated on custom-built phantoms. To foster a combined analysis and allow data harmonisation, additionally to the phantoms, healthy volunteers were scanned at each site [28]. Less targeted towards any specific disease cohort, is the UK Biobank[8] project. To date it is one of the largest, highly coordinated imaging studies planning to scan 100,000 volunteers[9] to collect data from 22 centres across the UK [175, 192]. The aim is to build a homogeneous database that, in combination with follow-up metrics (both clinical and imaging), can serve as an analysis pool for early biomarker detection within a middle-aged[10] population cohort. Although subjects appear to be healthy during the recruitment, patient groups will be identified with follow-up clinical assessments over time. It is estimated that in the sub-cohort that underwent MRI approximately 1,800 individual will have developed

[6] `www.tracktbi.ucsf.edu`
[7] `www.center-tbi.eu`
[8] `www.ukbiobank.ac.uk`
[9] as a subset of 500,000 participants, for which clinical variables were collected
[10] 40 to 69 years

Alzheimer's disease, 8,000 will suffer from diabetes and another 1,800 will have experienced a stroke [4, 175].

1.3 Challenges

1.3.1 Data Management

As previously established, neuroimaging databases are growing due to increasing number of subjects, enhanced image quality as well as multi-parametric and longitudinal image acquisition. This is fostered by breakthroughs in computational science (e.g. deep learning), clinical needs to study complex diseases in larger cohorts and improved infrastructure to acquire and share data. While beneficial for research purposes, storage, computing memory and computational speed become a difficulty for these large-scale databases. An essential part of neuroimaging studies is the removal of artefacts and the extraction of structured information relevant to the research question. Although this extraction usually condenses the available information and in theory compresses the database size, image processing tools often create intermediate output files that further increase the demand for (temporary) disk space. Moreover, some image analyses, such as computing brain connectivity on a voxel level from a functional MRI time series, can even lead to expansion of data size [233]. A centralised database that provides the same curated data for all researchers involved can help to minimised storage needs. This, however, requires an easily accessible infrastructure to share data, transparent *quality control* (QC) mechanisms and constant database maintenance [200]. With more imaging data available and the development of data-hungry algorithms, there is a shift towards use and optimisation of modern graphical processing units [152]. Besides advanced hardware and improved computing architecture, data management can further be supported by streamlined image processing pipelines. Since many software tools for neuroimaging data were not designed for large-scale analysis [252], there is also a need for new algorithms that accelerate processing performance. The combination of new state-of-the-art methods, integrated in optimised processing pipelines, and the design of modern computing clusters allow a high throughput of imaging data. Although optimal data management is the backbone to any research flow, this is only one of the challenges faced when dealing with large and heterogeneous databases.

1.3.2 Confounding Factors

Large neuroimaging studies can often be less specific as one of the primary goals is to amass a vast amount of data to cover the variability within a cohort. Population studies, such as for

example the UK Biobank project, have no single disease focus and data are collected before any subgroups are identified. Therefore, acquiring a big dataset is needed to improve the chances to detect subtle effects that were otherwise statistically not observable. Although large sample sizes are beneficial to highlight real effects, it also increases the sensitivity to confounding factors. This particularly holds true for large imaging studies, as the potential for imaging artefacts increases with sample size. Such artefacts can directly affect the imaging variables of interest. For example head motion or scanner hardware changes[11] are much more likely to occur in large and longitudinal studies [233]. Thus, it is important to apply sophisticated acquisition and processing methods to reduce the impact of these confounds while avoiding the alteration of the *true* signal [233]. For instance, accelerating image acquisition could help to reduce head motion artefacts. In addition, methods have been developed to reduce the effect of motion-related confounds, however, it was recently shown that this is non-trivial as different pre-processing steps can also influence subsequent data analysis [191]. On the other hand, there are confounding effects driven by the imaging hardware used and scanner software. These mostly will have the same effect on different subjects, however, will greatly vary for different imaging sites. For example, different vendors, magnetic field strengths and type of coils have been found to affect volumetric measurements extracted from anatomical MR scans [49, 120, 139]. Furthermore, the MR acquisition protocol was shown to impact image quality. For example, different magnitude [24] and number [41] of b-values were found responsible for variation in diffusion metrics. Both the angular [279] and spatial [194] resolution had an impact on the measured diffusion signal. Even when choosing identical imaging systems and setups, differences are inevitable, partially, since operator inconsistency can also alter image quality [233, 252]. Moreover, some confounding factors are unrelated to the image acquisition, but a result of anatomical differences between subjects. For example, cortical thinning is a natural age-related phenomenon and needs to be considered when examining tissue atrophy in diseased brains [233]. Other subjects dependent effects such as alcohol [243] and caffeine [148] consumption have been shown to influence MRI signal. Particularly when dealing with multi-centre studies, inequalities in databases need to be addressed to avoid any site-specific biases.

1.3.3 Data Harmonisation

Combining heterogeneous databases from multiple studies retrospectively often requires differences in image acquisition to be rectified. As a first step this could mean removing field inhomogeneities of individual scanners, resampling images to a common spatial resolution and/or projecting them to the same physical space. For prospective multi-centre studies,

[11] or inevitable hardware difference in multi-centre studies due to the different scanning location

with matched scanning parameters, site-specific biases are ideally less pronounced but cannot be ruled out completely. With more neuroimaging data available through collaborative projects there is a rising awareness of MRI inconsistencies across sites. A large empirical study showed that MR scans from 17 different databases could easily be associated with their origin by a standard machine learning classifier (≈72% accuracy) [258]. Similarly, it has been shown that even after meticulous pre-processing of anatomical brain scans, MR images from different studies could still be identified by a *randomised forest* (RF) classifier with high accuracy (≈80-99% for two databases) [81]. This indicates that commonly applied processing steps may not be sufficient in reducing site-specific differences. Even after applying harmonisation methods and successfully reducing inter-site variations, scans from different sites could still be distinguished easily (≈40-55% for 17 databases) [36, 258]. In fact, some techniques such as Z-scoring, were reported to also diminish the underling signal of interest [258]. This is especially problematic for prognostic analysis of multi-site data, where differentiation between patient cohorts could be hampered or even misled by site-specific influences. Therefore, standardised MR protocols have been employed to minimise differences across scanners [77, 92]. This is indeed a viable option for prospective multi-site imaging projects, however, currently many databases exists that could be harvested for clinical research if adequate harmonisation methods were available. Moreover, today's multi-centre studies may be tomorrow's legacy data. In other words, there will always be a trend to combine different databases to leverage the vast information from previously collected MR scans. Additionally, a recent survey has shown that more guidance is needed for medical data harmonisation [119]. In particular, the impact of current techniques on MR scans of different disease cohorts is little understood. For those reasons, harmonisation of neuroimage data remains an open research question that will become even more important with the increasing number of emerging multi-centre projects.

1.4 Overview of Thesis

Some of the core challenges for large-scale multi-centre studies involve optimising workflow for processing large-scale MRI databases, understanding the impact of different MR scanners and protocols on image derived features as well as general data harmonisation. This thesis aims to explore some of these concepts with direct application to a TBI cohort. Traumatic brain injuries are interesting to study, since diverse injury mechanisms and pathology development lead to highly heterogeneous disease patterns. While some patients show very subtle changes in WM, others experience severe head injury including lesions and brain tissue atrophy. This heterogeneity makes TBI a particular challenging cohort and some

concepts could also be transferred to simpler neurological diseases (e.g. tumour segmentation is less challenging than TBI lesion segmentation). At first, pre-processing pipelines will be introduced that were specifically, but not exclusively, designed for TBI data. These entail state-of-the art methods for artefact correction and preparing the data for subsequent analysis for clinical research questions (Chapter 2). Thereafter, the applicability of the pipelines will be tested on a mTBI study including data from three different sites. As a direct output from the pipelines, QC metrics will be employed to curate the dataset. Image derived features automatically computed via the pipelines will then be used for analysis of the mTBI cohort (Chapter 3). Furthermore, brain parcellation methods will be examined for their reproducibility when applied to multi-scanner data. Regional diffusion metrics will be compared for different scanners and/or MR protocols (Chapter 4). Acknowledging the site-specific differences in diffusion MRI, different harmonisation methods will be evaluated (Chapter 5). Finally, lesions of severe TBI patients scanned on two different scanners will be analysed (Chapter 6).

The following paragraph briefly describes the content of each chapter:

- Chapter 2 presents two flexible pipelines that allow processing anatomical and diffusion MRI. Besides the tailored artefact correction, the image derived features as well as the integrated QC are described.
- Chapter 3 showcases the application of the processing pipelines for TBI analysis. Regional volumetric and diffusion differences between controls and mTBI patients are examined and linked to patient outcome.
- Chapter 4 evaluates the robustness of anatomical white and grey matter parcellation methods with regards to different scanner hardware and investigates the influence of different acquisition protocols on diffusion MRI metrics.
- Chapter 5 compares different diffusion MRI harmonisation techniques and examines their application to a small mTBI study.
- Chapter 6 studies contusion and oedema in TBI patients scanned on two different scanners and introduces algorithms for automated lesion localisation as well as for lesion matching for longitudinal lesion analysis.
- Chapter 7 summarises the contributions of this thesis and provides an overview of future research directions.

Chapter 2

Data Analysis Pipelines

2.1 Requirements and Concept of Processing Pipelines

With data being collected for hundreds of subjects at several time points across different sites and vendors, datasets become quickly very complex [264]. As a consequence, a lot of time needs to be invested to curate and process a database such that it is ready to use for analysis by a multi-disciplinary team of clinicians, machine learning researchers and statisticians. Therefore, there is a demand for robust pipelines that facilitate image processing, data curation and information extraction. To build effective pipelines that support neuroscientific research they need to fulfil a number of requirements. To allow a high throughput of data, the pipelines need to be computationally efficient. This can be achieved by using fast methods, parallelising processes and connecting modules smartly together. Furthermore, data need to be processed in a robust way to minimise variation when rerunning the same pipeline on the same data. In addition, it is desirable to have a modular pipeline that allows the seamless integration of well-validated tools and new state-of-the-art methods. Since it becomes intractable to visually inspect large databases, QC metrics will need to assess the processed output and flag up corrupted scans and failed processing steps. Eventually, after removing artefacts and enhancing image quality, the pipelines will need to extract meaningful image features that can then be used to answer clinical questions. Ideally, input data fed to the pipeline and any processing steps are thoroughly documented, to provide information about what data were used and how it was handled.

Two pipelines were designed for this project, one for structural MRI, and another for diffusion MRI data. Both pipelines were implemented with the python library `nipype` (version

1.1.8) [86]. Alternatives could be FDT[1] (FSL's Diffusion Toolbox) or FreeSurfer [2], however, these mostly focus on algorithms developed for the particular toolbox. On the other hand, nipype is scriptable library facilitates the access to various sophisticated software packages (e.g. FSL, ANTS, MRtrix3, Dipy) via unified semantics. Many of them have been applied throughout multiple neuroimaging studies making results more replicable. One great advantage is, that once a processing module has been completed, it will be stored in a cache. So, when rerunning the pipeline on the same data again, processed data can be retrieved without any further computational effort as long as the input and the parameters for the particular module were unchanged. This makes the pipelines not only more efficient, but also more robust. Generally, if rerunning a processing module without any changes, the output quality might still be influenced by statistical variance of an applied tool (for example because of random initialisation of an algorithm). Besides the access to well known neuroimaging tools, nipype also enables the implementation of customised interfaces. This allows the integration of any command line tool or script, in order to link together the most suitable methods and tailor the pipelines to one's needs. The pipelines include modules to compute QC metrics for different tools as well as modules to extract *image derived phenotypes* (IDPs). Both QC metrics and IDPs are stored in *comma-separated values* (CSV) file format for each subject individually. After prepossessing a whole database, those CSV files can easily be concatenated within seconds to combine information for one study all together. Another great advantage of nipype is that all parameters for any applied processing tool are saved as text file such that the information can be retrieved and processes repeated any time. Furthermore, a custom text file is generated and stored for each processed scan, that exactly documents which pipeline version was run and what data were fed into it. The setup of both pipelines will be described in detail in the following sections.

2.2 Data Acquisition

Cam-CAN Database. Images for the Cam-CAN study were collected on a *three Tesla* (3T) Siemens Trio Tim scanner with a 32-channel head coil at the Medical Research Council (UK) Cognition and Brain Sciences Unit (MRC-CBU) in Cambridge, UK. The *T1-weighted* (T1w) *magnetisation-prepared rapid acquisition with gradient echo* (MPRAGE) scans were collected with an *echo time* $TE = 2.99\,ms$, a pulse *repetition time* $TR = 2250\,ms$, an *inversion time* $TI = 900\,ms$ and a 9° flip angle. A 2-fold acceleration factor was applied (generalised autocalibrating partial parallel acquisition) to cover the *field of view* (FOV) (256×240×192) with

[1] www.fsl.fmrib.ox.ac.uk/fsl/fslwiki/FDT
[2] www.surfer.nmr.mgh.harvard.edu

isotropic $1\,mm^3$ voxels. *Diffusion weighted images* (DWI) were acquired for two different b-values ($b = 1000, 2000\,s/mm^2$) with each 30 non-co-linear directions and three non-diffusion sensitised images ($b = 0\,s/mm^2$). Images with isotropic $2{\times}2{\times}2$ mm^3 voxels (axial slices; FOV = $192{\times}192$) were all acquired with $TR = 9100\,ms$ and $TE = 104\,ms$. More details are available in the Cam-CAN repository publication [239]. Although DWI are relatively low in angular resolution for today's standard, this represents a state-of-the-art imaging protocol given a limited acquisition time (approximately 10 minutes) at the time when the MR sequence was designed (multi-band imaging was not yet established).
To create a new age-unspecific T1w template and a corresponding atlas, T1w images of 652 healthy subjects (aged 18-87) from the Cam-CAN study were iteratively registered. For this, images were initially affinely registered to *Montreal Neurologcial Institute* (MNI) template space. After merging data to create an initial template (average image), images were repeatedly registered and merged, whereas for each new iteration the previous generated average image served as registration target. The iterative approach included three rigid, three affine and six deformable registration steps. T1-weighted images were parcellated in native space [151] (for more information see next section). Eventually, these individual parcellations were projected to template space with the corresponding transformation for non-linear registration (last step of registration iteration) and merged. All registration steps were performed with ANTS antsRegistration.

CENTER-TBI Database. For this multi-centre study multi-contrast MRI data were collected on 17 3T scanners (three different vendors: *General Electrics* [GE], Philips and Siemens) at 14 different sites. The various contrasts acquired during a scan session entailed T1w, *T2-weighted* (T2w), *fluid attenuated inversion recovery scans* (FLAIR) and *susceptibility weighted imaging* (SWI) as well as DWI.[3] Acquisition protocols were harmonised across all participating sites as well as possible, while coping with site-specific limitations (e.g. number of diffusion directions that can be acquired at once). Table 2.1 provides an example of MR parameters for one site (other sites are similar). Structural scans were acquired with $1{\times}1{\times}1$ mm^3 voxel size. Diffusion images were collected with $2{\times}2{\times}2$ mm^3 voxel size and the angular resolution varied between 30 or 32 gradient directions (site-dependent). Some centres collected a rescan of the same diffusion directions, which can be useful for analysing intra-scanner reproducibility. Further information can be found online.[4]

Scan-Rescan Database Twelve subjects (10 female, two male, age = 33.7 ± 6.8), were

[3] resting-state functional MRI was also acquired, but will not be specified here
[4] `www.center-tbi.eu/project/mri-study-protocols`

scanned twice within a few months on each of two scanners (Siemens Prisma & Trio) to acquire T1w MR image. In addition, six subjects underwent diffusion MR imaging during the same session. The scanning protocols were equal for T1w images except for slight difference in TE. A longer TE allows more time for the net magnetisation to return to its initial maximum value parallel directed to the main magnetic field. This results in decay of the T1w signal, which can lead to lower contrasts between tissues (given that the TR is short, and no T2w image is acquired, for which both TE and TR are long). A decreased contrast could directly influence image parcellation, however, this effect is expected to be minimal for this data due to the very small difference in TE (i.e. 0.04 ms). Besides the different TE, the total number of channel varied for the T1w images (40 and 32 for Prisma and Trio, respectively). Generally, a higher number of channels will result in a better *signal-to-noise ratio* (SNR), however, the spatial distribution of the receiver coils will need to be considered as well. Therefore, the impact of the different number of channels on this dataset cannot directly be assessed. Since the number of channel is fairly similar (only 8 more on Prisma than Trio), the differences are expected to be small. A recent study [199] has shown a high reliability between MR images collected on Prisma and Trio scanners, but a significant higher SNR for Prisma than Trio scans was found. This could be linked to larger (e.g hippocampus) or smaller (thalamus) volumes derived from Prisma scans (FreeSurfer). Diffusion imaging protocols mainly differed in the applied TR, but were otherwise mostly identically. On both scanners single-shell data ($b = 1000\, s/mm^2$) with 64 non-collinear gradient directions were collected. While for Trio DWI scans one non-diffusion sensitised image (b_0) was acquired, DWI scans on Prisma included five b_0 volumes. For better compatibility of both scanner protocols, the four additional baseline image were not considered. Since, these were acquired at the very end of the scan no artificial time gap between consecutive diffusion volumes was created. For both scanners 64 channels were employed. All scans were collected via parallel imaging in k-space (GRAPPA) covering 100% of FOV. Protocol parameters are summarised in Table 2.2.

2.3 Pipeline for Structural Magnetic Resonance Images

The pipeline for structural MR images takes several files as input. This consists of most importantly a T1w image and several other MR scans, here collectively defined as additional structural scans. Among those are T2w, FLAIR, SWI, *gradient echo* (GRE) and *proton density* (PD) images. The most commonly acquired MRI sequences are T1w and T2w scans, as they show excellent soft tissue contrast within the brain. While T1w images show brain anatomy most clearly (*cerebrospinal fluid* [CSF] dark and brain tissue bright), T2w scans can

Table 2.1: Example of Acquisition Parameters of CENTER-TBI Database

Contrast	TR	TE	TI	Flip Angle	PxBW
T1w	2300	2.98	900	9°	240
T2w	3000	222	-	90°	751
FLAIR	6000	394	2100	90°	781
SWI	29	20	-	15°	120
DWI	9800	91	-	90°	1698

TR: Repetition time in ms, TE: echo time in ms, TI: Inversion time in ms, PxBW: Pixel bandwidth in Hz/Px

Table 2.2: Acquisition Parameters of Scan-Rescan Database

	Prisma		**Trio**	
	T1w	DWI	T1w	DWI
FOV Read	256 mm	192 mm	256 mm	192 mm
Slice Thickness	1 mm	2 mm	1 mm	2 mm
TR	2250 ms	8500 ms	2250 ms	8400 ms
TE	3.02 ms	90.0 ms	2.98 ms	90ms
TI	900 ms	n/a	900 ms	n/a
Flip Angle	9°	90°	9°	90°
Pixel Bandwidth	230 Hz/Px	1628 Hz/Px	230 Hz/Px	1628 Hz/Px
Echo Spacing	n/a	0.77ms	n/a	0.72 ms
EPI Factor	n/a	96	n/a	96

also highlight pathology, which appears bright. Since, CSF also is bright on T2w images, FLAIR images are acquired which allows to suppress the signal originating from normal CSF (appears dark) while signal in abnormalities remain bright. Since all of the above mentioned scans are *three dimensional* (3D) images showing complementary information of brain anatomy, they are clearly differentiated from diffusion and functional MR images. Of course the latter show brain anatomy as well, but due to their more complex data structure, they will be processed in separate pipelines specifically designed for DWI. While the T1w scan is essential to run the structural MR pipeline, all other additional scans are optional and the pipeline would run completely without or with any subset of those scans. This is an important feature, as scan availability varies within and across studies, which may include different image contrasts. Usually, a T1w scan is always acquired by default and additional scans are collected as needed to answer research questions. But even scan sessions within

one study may be incomplete due to corrupted MR images or interrupted scanning session (for example due to patient discomfort). The flexible data input enables to automatically process different sets of anatomical scans more easily. Figure 2.1 provides an overview of the different pipeline modules.

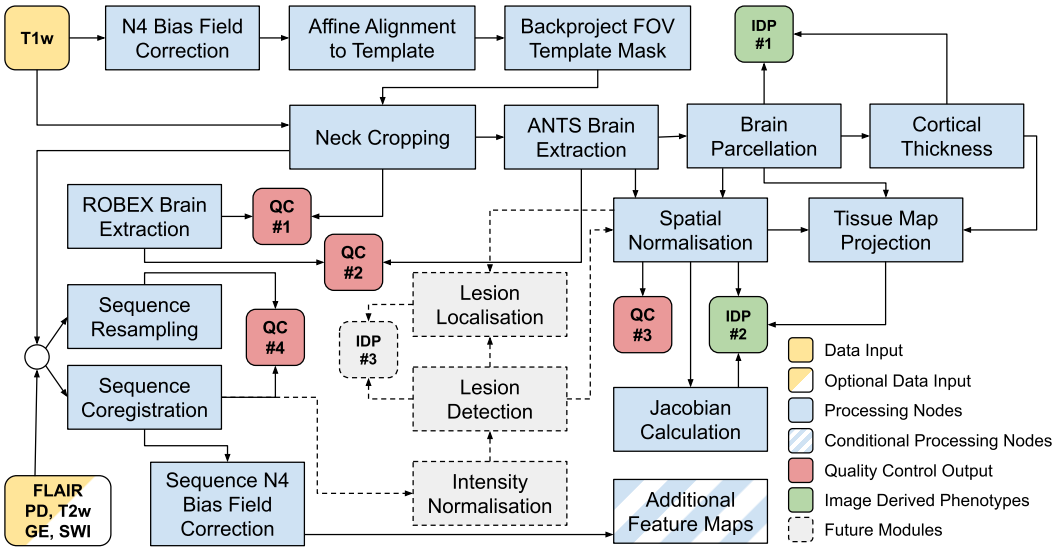


Figure 2.1: Schematic Overview of Pipeline for Structural MRI

2.3.1 Processing Modules

Neck Cropping. Although not the most pivotal step, cropping the neck reduces the number of voxels[5] and will accelerate many subsequent computing steps. Any registration based step, such as brain extraction and parcellation tools, will process images faster if less voxels are present. Besides reducing the computational costs, it also helps to make some processing steps more robust, as algorithms do not have to consider large areas solely including neck or background. In an earlier development stage, for example, it was observed that multi-atlas registration needed for the brain parcellation (see below) failed due to images showing a big proportion of the neck. So initially, the FSL [112] tool robustfov was used for neck cropping. However, this often led to cutting the image through the cerebellum and brainstem (Figure 2.2 B). This was mostly, but not exclusively, linked to images with tilted head position. Therefore, a simple but more robust tool was introduced (custom nipype interface): At first, T1w image intensities were corrected for inhomogeneity in the magnetic field, also known as N4 bias field correction [2]. This image was then affinely registered to a template (here: Cam-CAN template, Section 2.2) via ANTS antsRegistration [12]. Then, the transformation found was inversely applied (ANTS antsApplyTransforms) to the template's FOV mask

[5] voxels are the 3D building block of image volumes, analogous to pixels in a two dimensional image

(practically a mask that fills out the complete template image space). Subsequently, the most inferior point of the projected FOV mask was identified (if oriented in standard space, this is usually the mask's lowest z-coordinate, i.e. the lowest point along the neck) and used as location to crop the image within the transverse plane (Figure 2.2 C).

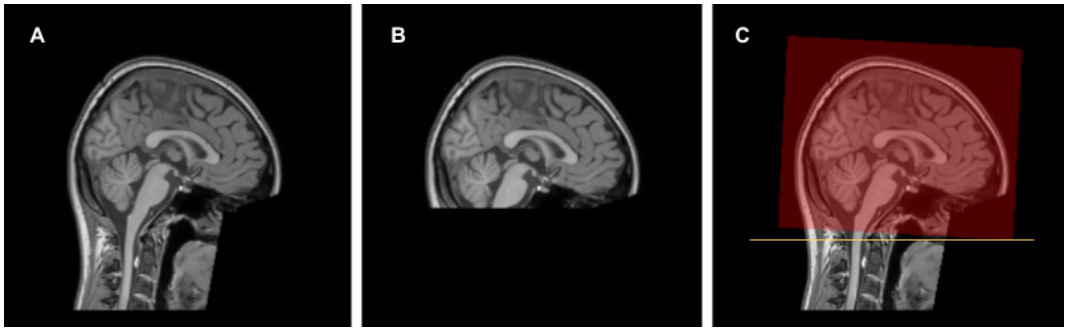


Figure 2.2: Example of Neck Cropping. As seen on the sagittal slice, the originally acquired T1w image included a big portion of the neck (A). Attempting to crop the image with common tool robustfov cut off parts of the cerebellum and brainstem (B). Affinely coregistering the T1w image to a template, and backprojecting the template's FOV mask (red area), allowed to find a more conservative lower bound (yellow line) to crop the image (C).

Eventually, the actual neck cropping is applied to the input T1w image, rather than the N4 bias corrected one, to preserve the original intensities. Applying robustfov to an early release of CENTER-TBI data (CENTER Core 1.0), including 1252 T1w scans, 132 scans (>10%) were incorrectly cropped. The customised approach turned out to be more conservative and has failed only in two cases (failed cases were identified as described later in Section 2.3.2 QC #1). These were then manually cropped to support affine coregistration to the template, before applying the same tool successfully again.

Brain Extraction. Generating a brain mask is often one of the first steps in brain image processing. This allows to focus the computational effort and the analysis only on voxels that actually show brain. Before integrating one or the other tool in the pipeline, several algorithms were applied to a severe TBI database (number of subjects N>100)[6], and compared directly against each other. Brain lesions and other abnormalities can drastically affect the quality of brain extraction, which is why the algorithms were tested on a clinical cohort. The included T1w images showed head deformations and lesions of various sizes and at different locations. After correcting T1w images for gradient field inhomogeneities (N4 correction), three different brain extraction tools were applied. The output of ANTS antsBrainExtraction [12], FSL's *brain extraction tool* (BET) [231] with different parameter settings and ROBEX [106] was blindly evaluated through visual inspection by a clinician. The total of 396 T1w

[6]This was part of a legacy dataset, as described in Section 6.2.1

scans included 136 scans of healthy controls as well as 260 scans of severe TBI patients. All three algorithms were designed for non-lesioned brains, which is why a perfect segmentation for all scans is unlikely. Hence, the superiority of an algorithms was defined as the one that failed on the least cases based on the visual quality control. Although BET extracted the brain accurately on 294 or 322 scans (depending on settings), it was outperformed by ROBEX and antsBrainExtraction with 348 and 353 successful segmentations, respectively. Furthermore, ANTS performed better than ROBEX on 18 scans. While BET often under- or oversegmented the brain, ROBEX and antsBrainExtraction performed almost equally well (Figure 2.3). In failed cases antsBrainExtraction included the eyes or CSF, which could be observed for both healthy subjects and patients. Since, antsBrainExtraction is based on a registration framework, the oversegmentation may be caused by inadequate alignment of T1w images to template space, however, no systematic bias was identified. Since subdural haematomas are found on the surface of the brain, antsBrainExtraction sometimes fails to include those lesions into the brain mask. This is patient specific and may be problematic if lesions were only located within the brain mask as subdural haematomas would not be detected.

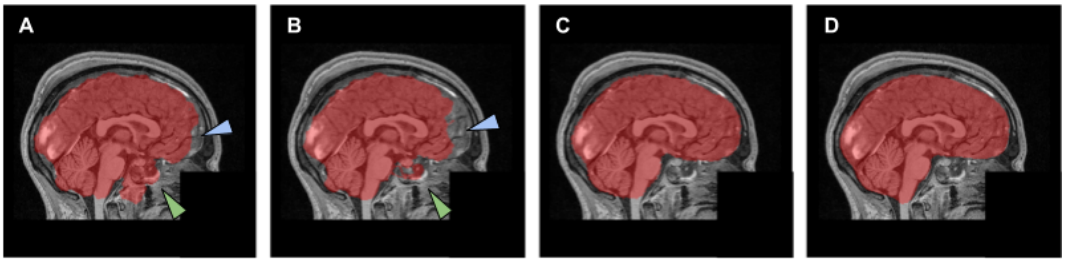


Figure 2.3: Example Comparison of Brain Masking Algorithms. Both options for BET with fractional intensity 0.4 (A) and 0.5 (default setting, B) showed over- (green arrow) and undersegmentation (blue arrow), here illustrated on a sagittal slice of a T1w image with frontal contusion. ROBEX (C) and antsBrainExtraction (D) performed equally well, both including the pathology. All masks highlighted in red.

The latter slightly tended to cope better with challenging anomalies in and around the brain. Skull stripping with ROBEX relies on brain shape priors, which cannot be changed, as they are part of the tool. Contrary, antsBrainExtraction is based on non-linear registration of the T1w scan to an atlas, and backprojection of a corresponding atlas mask from template to T1w image space. The atlas and its brain mask can be chosen by the user. Although being the slowest of all tools, antsBrainExtraction was eventually chosen for the pipeline, as it seemed to be most robust and allows for an external atlas to be provided, that might be best suited for challenging datasets (here: Cam-CAN template, Section 2.2).

Brain Parcellation and Cortical Thickness Estimation. One important aim in clin-

ical neuroscience is to examine whether for example a patient cohort differs from a control group. Usually, measuring global metrics within the entire brain, e.g. total brain volume or amount of *grey matter* (GM), is too crude to capture subtle differences and provides not enough information to draw a clinical conclusion. Therefore, a common process is to subdivide the brain into different anatomical regions, which allows for changes to be observed on a more local level. Once the brain is parcellated, information such as volume, GM content or cortical thickness, can be retrieved from each individual region. This enables a more detailed analysis while still summarising the images' information in contrast to considering all voxels as a single unit (e.g. voxel based morphometry).

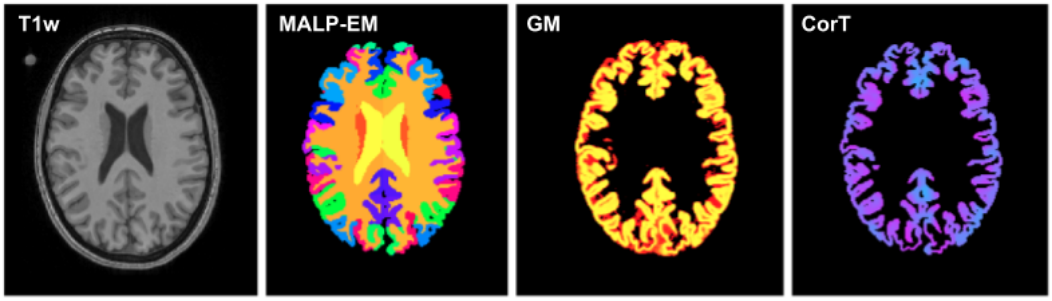


Figure 2.4: Example of T1w Image and Computed Feature Maps from a Healthy Subject. T1w bias-corrected image (T1w) and the corresponding brain parcellation (MALP-EM, random colours for different ROIs), its GM probability map (GM, yellow indicates a higher probability than red) and cortical thickness map (CorT, blue indicates a thicker cortex than purple).

The *multi-atlas label propagation with expectation maximisation based refinement* (MALP-EM) tool was chosen for parcellation, as it has been shown to work well for deformed brain anatomy including TBI cases [151]. The better performance of MALP-EM can be attributed to the multi-atlas segmentation approach and the retrospective intensity refinement based on the segmentation posteriors. In theory, any region atlas could be used in combination with MALP-EM's underlying algorithms, however, this would require several pairs of T1w images and their corresponding ROI atlases. In brief, this approach makes use of 30 atlases from the *Open Access Series of Imaging Studies* (OASIS) database [167] that have been manually subdivided into 138 *regions of interests* (ROIs). After non-linear registration of the 30 atlases to the T1w image that is supposed to be segmented, the manual label maps are projected onto the T1w image and merged to calculate probabilistic posteriors for each ROI. With those posteriors[7] the parcellation is refined via expectation maximisation. The registration is based on T1w image intensities as well as tissue segmentation maps, so that besides the brain parcellation MALP-EM provides also tissue probability and label maps.

[7] the posterior distributions are actually first relaxed to facilitate the refinement

The tool can easily be applied via the command line, but it was integrated in the nipype pipeline for a streamlined use. Although MALP-EM comes with an innate brain masking algorithm, the mask previously computed with ANTS was used, as it provided more accurate and more robust results. The output of probabilistic tissue maps for GM and WM were further fed to the *diffeomorphic registration-based cortical thickness* (DiReCT) [46] estimation (via nipype's built-in interface for the ANTS KellyKapowskialgorithm). Figure 2.4 shows an example of T1w images and the corresponding maps for MALP-EM parcellation, GM segmentation and cortical thickness estimation.
One of the IDPs that can be computed from the brain parcellation are the ROI volumes. Besides the 138 ROI volumes, MALP-EM further provides summarised volumes for total brain volume, cortical as well as deep GM, WM and CSF. Furthermore, the average tissue densities or cortical thickness (i.e. the mean value of the GM probabilities within a region) could be computed within each ROI (IDP #1).

Processing of Additional Magnetic Resonance Sequences. The strength of multiparametric analysis lies in observing several MR contrasts simultaneously to derive clinically useful information. To achieve direct correspondence between all MR scans for the same individual, the different image contrasts were aligned to one another, commonly known as coregistration. Usually, a high resolution (1mm^3) T1w image is acquired by default, which is why it served as reference to which all other images were registered to. Therefore, ANTS antsRegistration was integrated to rigidly[8] align all additional MR sequences to the neck cropped T1w image. *Mutual information* (MI) served as similarity metric during optimisation. All sequences are treated equally and individually, so that regardless of which sequence was present, the pipeline will align all available raw images. However, since T2w and PD were acquired simultaneously, the pipeline was designed to keep the perfect alignment between PD an T2w images intact. So, in case both PD and T2w images were present the transformation of the T2w to T1w scan was simply applied to the PD image (antsApplyTransforms). Furthermore, all additional scans were also N4 bias corrected, analogous to the T1w image. After that, the corresponding transformation for each scan was applied to the bias corrected image. This order was chosen, to prevent any possibly failed N4-bias correction (for example because of pathology) to hamper the coregistration process. Eventually, all raw and bias-corrected images were projected to T1w image space.
After coregistration two additional feature maps were computed, given that the particular scans were available: One was the ratio of the aligned bias-corrected T1w and T2w image (T1w/T2w), because this has been suggested as a surrogate for mapping myelin. It has been

[8] only translation and rotation of moving image

shown that signal intensities on those *myelin maps* correlated well with subcortical myelin development [74]. Measuring a surrogate of the myelin content could help to estimate WM related brain atrophy if diffusion weighted MRI was unavailable. The other feature map was the product of the aligned bias-corrected T2w and FLAIR image ($FLAIR^2$), which has been shown to provide higher grey-white matter *contrast-to-nose ratio* (CNR) than the standard FLAIR image. Furthermore, it has been observed to yield an increased CNR between WM and lesions in comparison to T2w and FLAIR images [266]. Since it has been reported to improve automated segmentation for multiple sclerosis lesions [149], this enhanced image contrast could also be beneficial for analysing other WM pathologies.

Spatial Normalisation. The spatial normalisation was performed with ANTS as it has been shown to be one of the most capable algorithms [134]. The process was split into two stages: the linear and the non-linear registration of the T1w scan to the template (antsRegistration). A custom template created from the Cam-CAN T1w scans (Section 2.2) was chosen over the standard MNI atlas. The former provides better anatomical detail and is less age biased, due to the larger number and wider demographic range of included subjects. The linear registration entails both the rigid and affine transformations. At first, the rigid registration phase aims to align both images via translation and rotation of the T1w image. Thereafter, the similarity function is further optimised in the affine registration stage by allowing shearing and scaling alongside rigid transformations. For both stages MI between the two whole images served as cost function for optimisation. The non-linear registration (SyN) was then initialised with the previously found affine transformation. For this stage, cross-correlation within the brain mask was chosen as similarity measurement.
As previously mentioned, the integrated brain extraction was also based on spatial normalisation. To understand whether the additional spatial normalisation provided any benefit over the deformable alignment entailed in the brain extraction (both are based on ANTS SyN), their effect on 424 T1w scans from the severe TBI database was compared. For this, the transformations found in both approaches were first equally applied to the same T1w images. Then, the *normalised cross-correlation* (NCC, also see Equation 2.2) between all pairs of warped images and Cam-CAN template was computed. In 98% of all cases (414 out of 424) the additional deformable registration resulted in a higher NCC than the initial spatial normalisation during brain extraction. This can likely be attributed to a longer optimisation phase and uni-modal registration approach. The optimisation process focuses on minimising the cost-function between template and T1w image only. In contrast, the spatial normalisation performed during brain masking is driven by multi-modal registration, including T1w images and tissue prior maps, which may hamper the optimal alignment.

In addition, a more customised spatial normalisation could also be beneficial when working with brain scans that show lesions (see below).

White and grey matter probability maps are generated as a by-product of MALP-EM ROI parcellation. Both tissue maps and the maps for cortical thickness are projected to Cam-CAN atlas space. By calculating the gradients at each element of the deformable transformation field, a Jacobian matrix was derived (ANTS CreateJacobianDeterminantImage), indicating the relative position of neighbouring voxels. The Jacobian determinants are often used to modulate the intensities of the spatially normalised tissue maps, to preserve the actual amount of tissue within each voxel (multiplication of Jacobian with tissue probability maps after spatial normalisation). Besides that, they also have been shown to be useful for analysing brain volume changes in moderate to severe TBI patients [38]. If satisfied with the spatial normalisation and using a study-unspecific atlas, the user could compare the maps for tissue density, cortical thickness and Jacobian determinants in a *voxel-based morphometry* (VBM) [10] analysis right after the data have been fully processed (IDP #2).

Lesion Data Processing. Although this branch of the pipeline is not yet fully integrated, it will still be described to explore ideas about how modules could be tied into the grand scheme of the structural MRI pipeline.

A very important first step is the detection and segmentation of a lesion. For the purpose of a fully automated pipeline, a lesion segmentation tool is needed, that does not require any user interaction. Current state-of-the-art methods for lesion segmentation involve *convolutional neural networks* (CNNs). For the presented pipeline, a version of DeepMedic [125] will be integrated, as it has been shown to outperform many other approaches (including other neural networks) for different types of lesions such as TBI [124] or stroke [123]. DeepMedic is a supervised neural network, meaning it is trained on MR scans and their corresponding label maps, which specify the presence of a lesion on a voxel-wise level. Besides differentiating between lesion and healthy tissue (binary case), DeepMedic can also learn to distinguish lesion types (multi-class case). It can be trained on single image contrast, but performs best when multiple (ideally complementary) MR contrast are used. Furthermore, this open software tool has been developed by collaborators on the CENTER-TBI project, so that their expertise can be directly leveraged for the pipeline. Usually, before feeding the images to a neural network, their intensities are standardised. One approach would be to calculate global mean (μ) and standard deviation (σ) of the voxel intensities within a brain mask for each scan individually, and then subtract μ and divide by σ (Z-score normalisation). To deal with gross intensity outliers, the mean and standard deviation can be computed only for voxels within a percentile range. This process results in centred image intensities

(μ=0) with unit variance (σ^2=1) [109]. This potentially increases comparability of images from different scanners an sites (multi-centre data), but also supports the training of CNNs. Features of similar range help numerical optimisation during neural network training.
From the segmented lesion, it is straightforward to compute total and relative lesion volume (IDP #3), however, it would be also important to analyse where the lesion is located within the brain. This could be achieved by comparing the detected lesion against an atlas and derive informative metrics. More details for brain lesion localisation can be found in Chapter 6. Once a lesion is automatically detected, it could also be used to improve the spatial normalisation, since cost function masking of lesioned areas has been shown to be necessary for deformable registration of brains with large lesions [5, 27].

2.3.2 Quality Control Metrics

The processing tools were selected and designed to be as robust as possible, however, it is important to check the quality of their output. Quality control metrics should help to highlight situations where the processing steps failed or corrupted the data. Good QC metrics are robust and provide an interpretable, quantitative measurement that can be computed efficiently.

QC #1: Neck Cropping. First, the ROBEX mask was computed on the raw T1w image. After cropping the T1w image, the cropping mask found was applied to the ROBEX brain mask as well. Since, the aim is only to prune the neck, the brain volume estimated by the ROBEX mask should be unaffected. This concept was tested on the CENTER-TBI database. If there was a discrepancy between brain volumes before and after neck cropping, the images were checked visually. Only two out of 1245 cases failed, which were manually cropped with fslroi and then pre-processed successfully.

QC #2: Brain Masking. The earlier mentioned experiment for brain masking has shown that antsBrainExtraction and ROBEX were both fairly robust. Ideally, they would generate the same brain mask resulting in equal brain volumes. In practice this is rarely the case due to the different nature of the underlying algorithms. Nonetheless, if both worked accurately the ratio of the extracted brain volumes should be close to 100% and any strong deviation might flag a sup-optimal performance of one of the two algorithms. When plotting the ratio of the total brain volume as extracted from antsBrainExtraction (V_{ANTS}) and ROBEX (V_{ROBEX}) for all CENTER-TBI T1w scans (139 control & 1252 patient scans), the QC metrics showed a slightly different distribution for controls and patients (Figure 2.5 left). While the average ratios were fairly similar for controls (104.8%) and patients

(103.8%), there was a large number of patient scans with brain volume ratios below 98%. Both subject categories tended to have a larger ANTS than ROBEX brain volume, which was cohesive with the general observation when checking brain masks visually. Inspecting both masks for outliers revealed different causes for the diverging brain volumes. The first scan (Figure 2.5 A1 & B1) displays the control subject with the highest volume ratio (111.7%), which resulted from ROBEX under segmenting the cerebellum. The outlier with the highest ratio (128.7%) showed that ANTS strongly oversegmented a patient's brain laterally (Figure 2.5 A2 & B2). The large amount of soft tissue around the skull, might have hindered an accurate spatial normalisation during antsBrainExtraction. At the lower end with ratios lower than 95% it was observed that ROBEX commonly oversegmented patients' scans as displayed in Figure 2.5 A3 & B3 (90.0%). Furthermore, ROBEX was found to underperform in the presence of lesions. It seemed to exclude pathological areas which consequently led to undersegmentation (111.8%, Figure 2.5 A4 & B4)

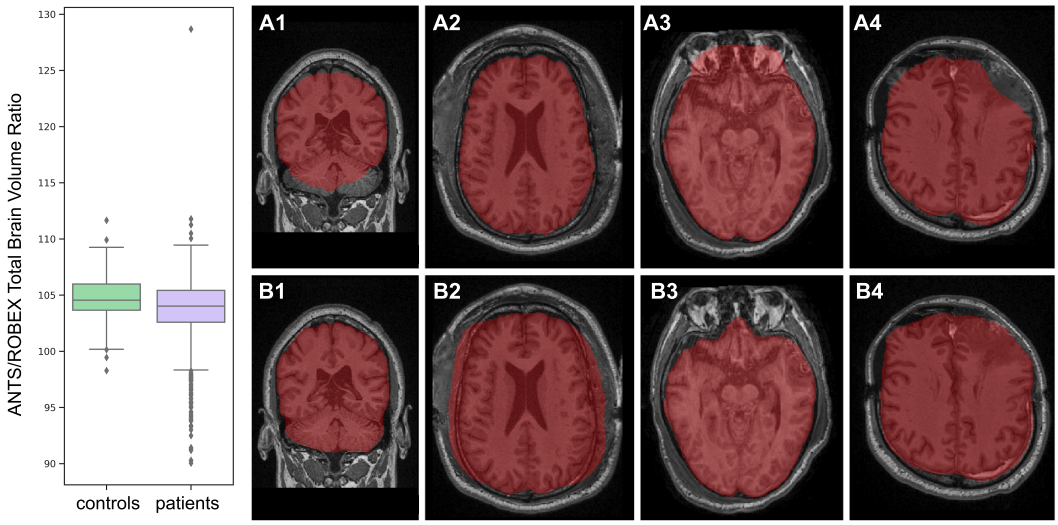


Figure 2.5: Quality Control of Brain Extraction for CENTER-TBI Database. **Left:** The distribution of the brain volume ratio shows wider spread for patients (Inter Quartile Range: IQR=2.8%) than for controls (IQR=2.3%). **Right:** Top row shows the ROBEX brain mask on four different scans with different performances: Undersegmentation of cerebellum in controls scan (A1), adequate brain extraction for patient (A2), oversegmentation of patients frontal brain region (A3), exclusion of lesion tissue (A4). Bottom row shows antsBrainExtraction performance on the same scans: It failed to deal with large amount of tissue surrounding the skull (B2), but performed better than ROBEX for all other scans (B1, B3, B4).

QC #3: Spatial Normalisation. Experiments that require spatial correspondence between different subjects or alignment to a template (e.g. VBM or lesion localisation) rely an accurate registration to a template. When correctly spatially normalised, the template and warped image would show similar anatomical structures at certain locations. This similarity was estimated by computing the NCC between both images images. The closer the value is

to one the better the image matching. The NCC was computed within the template brain mask as follows:

$$NCC = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x\,\sigma_y} \tag{2.1}$$

where x and y represent both images and mean (μ_x) and standard deviation (σ_x) were defined as (analogous for y):

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2} \tag{2.2}$$

Considering such a QC measurement before any analysis could help to spot failed spatial normalisation and flag corrupted scans, that both, if included, could adversely influence the final results. Extracting this information from the QC report from the structural pipeline for all CENTER-TBI T1w scans, a clear difference between scans from controls (mean=0.804) and patients (mean=0.778) was observed. While NCCs for controls stayed well above 0.70, NCCs for patients were as low as 0.37.

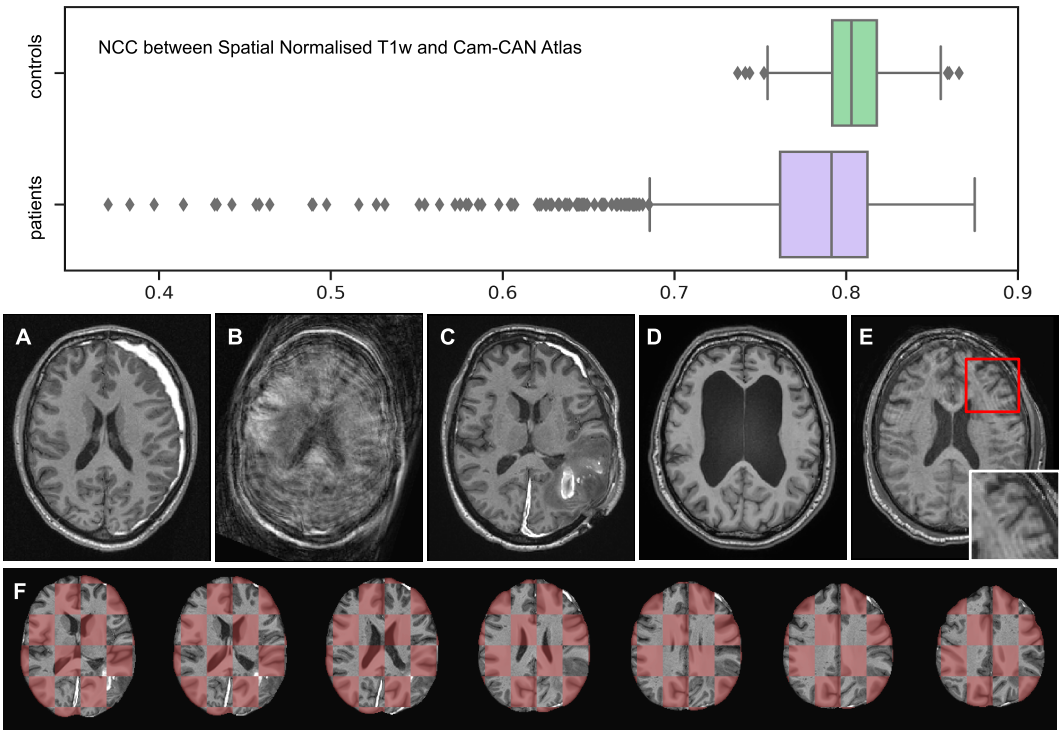


Figure 2.6: Quality Control of Spatial Normalisation for CENTER-TBI Database. **Top:** Scans for patients had much more distributed NCCs (*interquartile range* IQR=0.051) than controls (IQR=0.026). **Middle:** Scans with low NCCs exhibited prominent image features such as lesions (A, C), head motion artefacts of different severity (B, E) or tissue atrophy and enlarged ventricles (D). **Bottom:** The mosaic, chequerboard images alternating between atlas and registered images provide the means for a quick visual QC (F).

Figure 2.6 shows five examples of spatially normalised T1w scans from patients with some

of the lowest NCCs. The first example shows a brain scan with large subdural haematoma (Figure 2.6 A). Its low NCC (0.370) could be explained by the high intensities within the lesion, skewing the measurement unfavourably. Although the spatial normalisation was not necessarily unsuccessful, such a case probably should be excluded in a voxel-wise group analysis and for any parcellation based extraction. To derive a NCC that is less sensitive to pathology, the similarity metric could be computed in healthy tissue only. This, however, would require a lesion annotation (manually or automatically generated) and in the presence of large lesions the similarity metric would be computed on a lower number samples, which may hamper the comparability to other scans. Another case with low NCC (0.383) displayed a T1w scan severely corrupted by motion (Figure 2.6 B). Scans with severe blurring must be excluded from any analysis. A low NCC could also indicate the presence of dominant anatomical abnormalities, such as a large contusion and oedema (NCC=0.397, Figure 2.6 C) or significantly enlarged ventricles (NCC=0.432, Figure 2.6 D). Severe pathological patterns can impede spatial normalisation and should be excluded from analysis reliant on spatial correspondence. Besides flagging up cases with severe deviations from the template due to artefacts or pathology, NCC was also found to gradually change with scan quality. A moderate low NCC (0.605) could highlight less dominant acquisition artefacts, such as ringing from head motion (Figure 2.6 E). Those scans might still be used for certain experiments, but have to be inspected to avoid potential biases in subsequent processing steps or analysis. By design, the pipeline also provides a mosaic overview of the chequerboard image, alternating between the template (here Cam-CAN in red) and the spatial normalised T1w image (grey, excerpt shown in Figure 2.6 F). This allows a quick and easy visual quality check. The chequerboard displayed here corresponds to the scan in Figure 2.6 C. Although the NCC was quite low, the registration seemed successful, as boundaries of anatomical structures (see ventricles) seamlessly continue across the mosaic tiles. The findings suggest that low NCC could pick up on a variety of different problems for spatial normalisation, including image artefacts and brain pathologies.

QC #4: Coregistration. Multi-parametric analysis requires the alignment of different sequences from one scan session. For example, coregistered scans are needed for predictive lesion segmentation, or if anatomical information such as ROI parcellation should be used to derive features from other sequences than T1w images. Similarly to the spatial normalisation, the accuracy of coregistration can be estimated by computing the similarity between aligned images. Again NCC was chosen as it is less sensitive to intensity scaling than, for example, mean absolute error and can be applied to sequences with different contrast. The metrics were calculated between the T1w image and a resampled image of the sequence

(matched resolution but no change of position or orientation) as well as between the T1w image and the coregistered scan. If successful, the metric should be closer to one after coregistration than before. The quality assessment on the CENTER-TBI database was two fold. Firstly, to flag scans that had a generally low NCC, and secondly, to inspect scans with a lower NCC after coregistration than before.

Table 2.3: NCC for Coregistered Sequences. Values displayed as mean [IQR].

	Controls		**Patients**	
	resampled	coregistered	resampled	coregistered
T2w	0.364 [0.171]	0.696 [0.094]	0.353 [0.145]	0.666 [0.116]
FLAIR	0.393 [0.499]	0.830 [0.096]	0.376 [0.186]	0.796 [0.146]
SWI	0.220 [0.201]	0.632 [0.106]	0.207 [0.179]	0.663 [0.102]

Besides T1w images, T2w, FLAIR and SWI images were available for the CENTER-TBI database. Table 2.3 lists the average NCC results for the resampled and coregistered sequences for both healthy controls and patients. The improvement was quantified by the ratio of average NCC of coregistered over resampled scans ($NCC_{coregistered}/NCC_{resampled}$ $\times$ 100%). The NCC improved for all T2w (average NCC ratio = 191%) and FLAIR (average NCC ratio = 211%) scans of the control group through coregistration. Only three SWI controls scans did not improve upon coregistration, however, this was due to severely corrupted images (Figure 2.7 A), and overall the coregistration was advantageous (average NCC ratio = 287%). For patients the NCC was higher after registration for all 1211 FLAIR scans (average NCC ratio = 212%). Just eight out of 1251 T2w scans and 14 out of 1230 SWI scans for the patients showed a lower NCC after coregistration. Nonetheless, there was an average overall improvement for T2w and SWI patient scans of 189% and 321%, respectively. The best relative improvement was yielded for SWI patient and control scans, which is likely due to the low starting point ($\approx$ 0.2) to begin with. Furthermore, flagged scans could be associated with poor image quality (Figure 2.7 A, B and C1). Interestingly, the NCC for patient scans was always slightly lower than the NCC for control scans, except for coregistered SWI scans. This may be explained by the presence of lesions that are more prevalent on MR contrasts other than T1w images. Moreover, this could indicate a bias in the acquisition of patient scans, possibly caused by increased head movement resulting motion-induced blur, ultimately leading to lower image similarities.

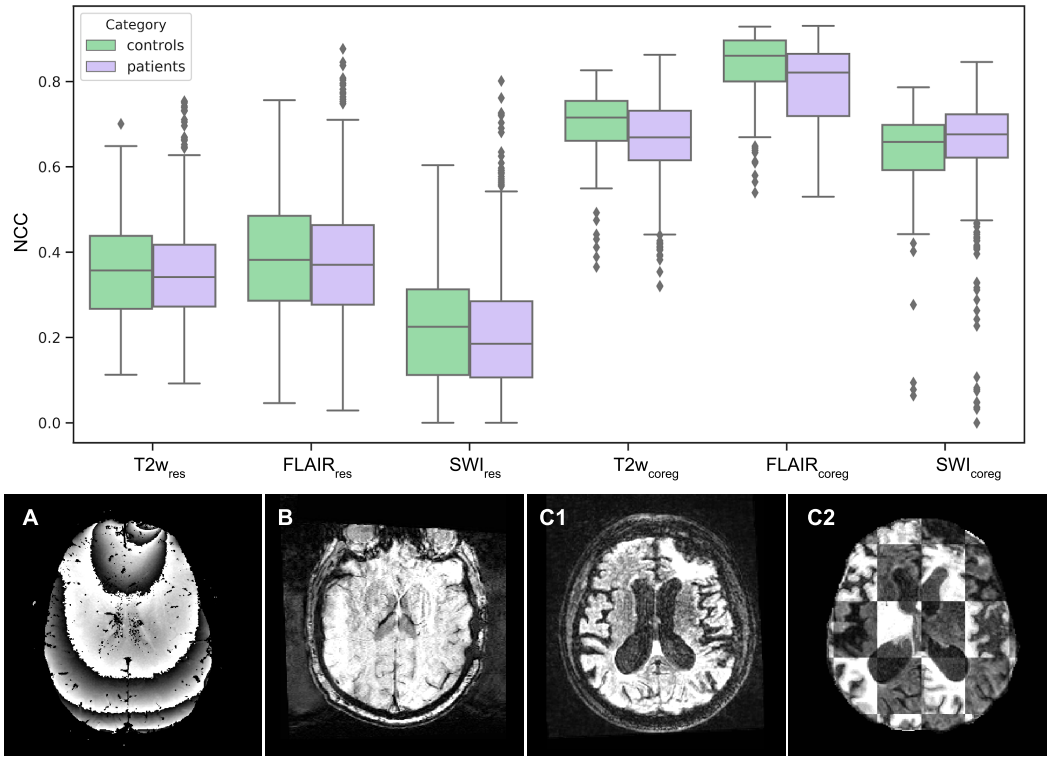


Figure 2.7: Quality Control of Coregistration of CENTER-TBI Database. **Top:** NCC between T1w and addition sequences improved for all three contrasts through coregistration ($T2w_{coreg}$, $FLAIR_{coreg}$, SWI_{coreg}) in comparison to simple image resampling ($T2w_{res}$, $FLAIR_{res}$, SWI_{res}). Mostly, NCC was slightly higher for controls than for patients, however, both groups were comparable for all three sequences before and after coregistration. **Bottom:** Scans with low NCC could be associated with poor image quality, such as corrupted acquisition (A) and noise artefacts (B) on SWI scans. A noisy FLAIR scan was flagged with a low NCC (C1), however, structures between FLAIR and T1w images still aligned as seen in the chequerboard image (C2).

2.4 Pipeline for Diffusion Magnetic Resonance Images

The pipeline for diffusion MR images also takes several files as input. Most importantly this is the diffusion weighted images (DWI) and its corresponding text files with the information for b-values (bvals) and b-vectors (bvecs). During DWI, different gradients of magnetisation are applied to encode a spatial correspondence of received MR signal. While the duration and strength of the employed gradients are defined by the b-values, the direction is indicated by the b-vectors. Diffusion along the b-vectors direction causes signal attenuation which is reflected in lower intensities in the diffusion sensitised scan. The diffusion pipeline takes as further input a T1w image and its associated brain mask. The pipeline coregisters *diffusion tensor imaging* (DTI) parameter maps to the T1w image (see below). Thus, for optimal results the T1w image must be from the same subject as the diffusion images and ideally also from the exact same scan session. Furthermore, it is advisable to use the cropped

and bias-corrected output files from the previously described structural MRI pipeline. After applying both pipelines, all anatomical MR sequences and the diffusion feature maps for one scanning session are aligned all together in the T1w image space. For artefact correction the pipeline also expects a text file with acquisition parameters (ACQP). In addition, the user has the option to provide a previously computed or manually delineated DTI mask and an extra b_0 volume. Both options will be discussed in more details below. Figure 2.8 provides an overview of the different components of the pipeline.

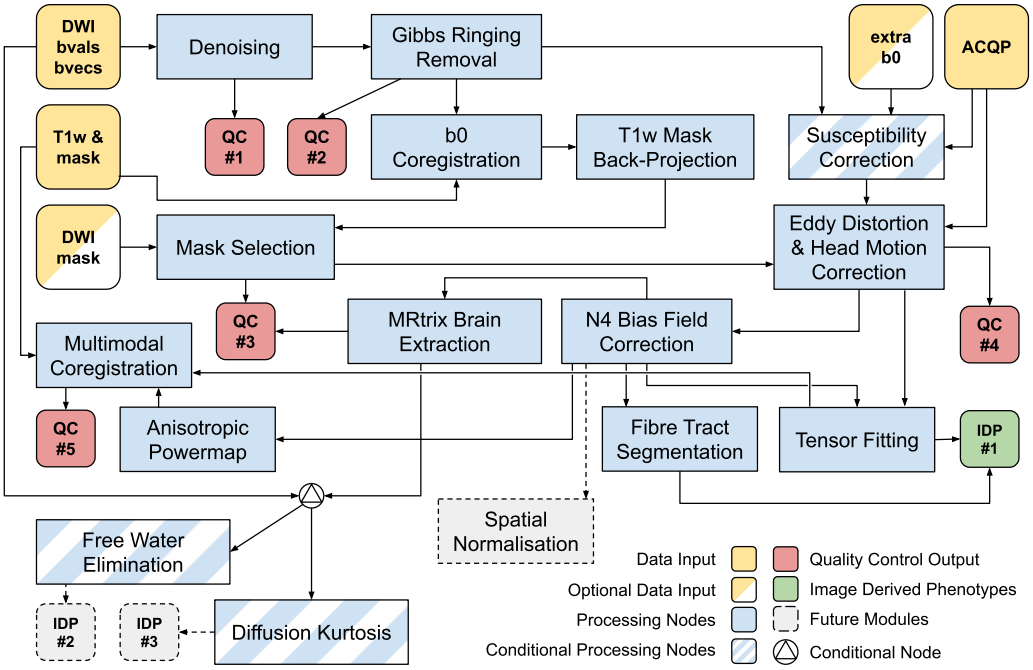


Figure 2.8: Overview of Pipeline for Diffusion MRI

2.4.1 Processing Modules

Denoising. Diffusion weighted imaging is affected by acquisition noise stemming from signal reconstruction. Especially for images with low SNR, the noise distribution deviates from a Normal distribution [90, 55], which makes noise reduction non-trivial. This is particularly important for advanced DWI that includes higher b-values, as this leads to images with lower SNR, which could hamper the correct estimation of diffusion measures [197, 118]. One effective way of suppressing noise is based on local *principal component analysis* (PCA). Assuming multi-directional signal redundancy within the diffusion images, PCA can be used to find components that are associated with statistical noise rather than structured information. By excluding components of low magnitude noise can be reduced. The initial suggestion to eliminate components by thresholding the associated eigenvalues [166] was

later improved through a data driven way through matrix theory (*Marcenko-Pastur-PCA*, MPPCA) [255]. This technique has been found to reduce noise while preserving anatomical structure and outperformed other approaches, such as adaptive non-local means or total generalised variation. Denosing is the very first step for image enhancement in the diffusion MRI pipeline, and was integrated via nipype's MRtrix3 interface for dwidenoise. Figure 2.9 shows an example of a multi-shell DWI image before and after denoising. The difference between both images shows that mostly random noise was eliminated.

Gibbs Ringing Removal. Magnetic resonance images are reconstructed by applying Fourier transforms to the raw signal. In practice, only a finite number of frequencies can be sampled, which is why the images are only approximated by a few components of the Fourier representation. This signal truncation is unproblematic for gradual changes in signal intensity (i.e. homogeneous regions), however, results in multiple fine ripples directly adjacent to high-contrast interfaces (e.g. the boundary of brain tissue such as at ventricles). Also known as Gibbs ringing, this artefact can be diminished by increasing the number of phase-encoding steps [9] or reducing the FOV, but cannot be completely avoided [44, 254]. A variety of methods have been suggested to reduce this effect of signal truncation. For the pipeline the MRtrix3 command mrdegibbs was integrated via a customised interface.[10] This algorithm is based on the technique proposed by Kellner et al. [130]. Here, the idea is to sample the image not at the sinc-function's extrema but rather close to its zero-crossings. To achieve this for all edges across the image, a local subvoxel-shift is enforced.

Brain Extraction. Many tools for brain extraction of diffusion MR data rely on thresholding the b_0 volume, as this usually provides the highest SNR and highest contrast between brain tissue and background. However, thresholding is not very robust and error-prone for noisy data. For the pipeline a simple method based on coregistration to T1w images was integrated. Therefore, the denoised and Gibbs artefact corrected b_0 volume was rigidly coregistered (antsRegistration) to the provided input T1w image. Subsequently, the input mask for the T1w image was backprojected (antsApplyTransforms) to diffusion image space. This approach underlies two assumptions: First, the brain mask for the T1w image must be accurate. Second, both T1w and b_0 volumes show the same brain shape and size, and neglects distortions resulting from *echo-planar imaging* (EPI) of the diffusion images. The user has also the choice to provide a DWI brain mask, in which case the automated brain masking would be suspended to save computing time. The pipeline recognises which path-

[9] usually less samples are collected in phase-encoding direction making the artefact more noticeable
[10] mrdegibbs is part of nipype since version 1.2.0, thus the custom interface will be replaced in future

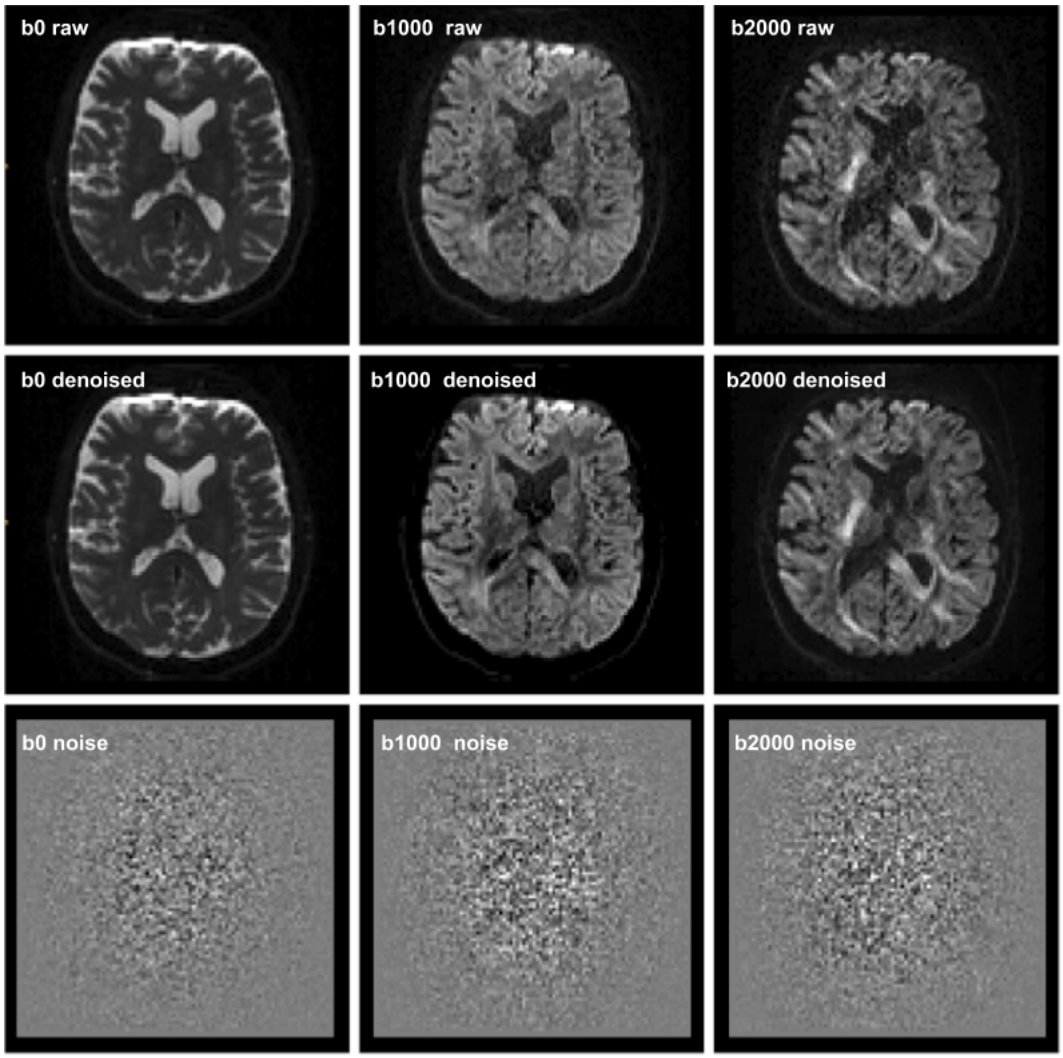


Figure 2.9: Example of Multi-Shell DWI Denoising from Cam-CAN Database. The first row shows a representative images for all three different b-values: Left to right, the diffusion unsensitised image, the first shell with b=1000 s/mm^2 and the outer shell with b=2000 s/mm^2. The second and third row show the corresponding images after MPPCA denosing and noise (difference) maps, respectively. The average absolute difference within the brain mask of the noise was 2.9 for the b_0 volume and 3.7 for both diffusion sensitised volumes. Image contrasts were adjusted for each shell individually, noise was scaled equally for all three difference maps for direct comparison.

way - using the input or automatically computed mask - has been triggered and selects the appropriate mask for the subsequent processing steps (Figure 2.8: Mask Selection). After head-motion correction (see next paragraph) another brain mask was calculated with MRtrix3 dwi2mask [53] mostly for subsequent QC (Section 2.4.2 QC#3).

Distortion & Head-Motion Correction. When performing spin-echo EPI, the resulting diffusion images can look distorted due to the acquisition's sensitivity to non-zero off-resonance fields. These will be induced by both the susceptibility distribution of the subject's head and eddy currents resulting from rapid changes of the diffusion weighted gradients. Additionally, for DTI multiple images with different gradient directions need to be acquired, which leads to longer scanning times and unavoidable head motion. Strong susceptibility induced artefacts can be problematic when comparing to undistorted T1w images. Furthermore, the distortion of anatomy might introduce inaccuracies when extracting region-based IDPs. If an extra b_0 volume was collected with reversed polarity of the phase-encoding blip relative to the diffusion weighted images, this can be used to correct for susceptibility artefacts (b_0 field map correction is not yet supported). Providing the images with opposing phase direction and the acquisition parameters, the pipeline estimates the susceptibility induced off-resonance field via FSL's topup [6]. This tool aims to find the deformation field that would match the b_0 volumes with opposing directions of distortions such that their similarity (sum-of-squared differences) is maximised. Once the susceptibility induced field was approximated, topup output can be fed to FSL's tool eddy , which concurrently corrects for eddy current distortions and head movements [7, 8]. In case no extra b_0 volume was acquired, the pipeline skips the correction of susceptibility distortions and moves straight to reducing eddy current and head motion artefacts. The pipeline currently does not support any alternative correction of susceptibility artefacts.

Bias Correction. To remove effects of the inhomogeneous magnetic field, a bias correction approach was included in the pipeline. This followed the approach as initially suggested by Jeurissen et al. [114]. First all non-diffusion weighted images (b=0 s/mm^2) within the scan are averaged before estimating field inhomogeneity via the N4 algorithm (analogous to T1w images). Subsequently, the bias is removed from all diffusion images separately. Since this approach assumes spatial correspondence between all DTI volumes, this process is applied after the head motion correction. This method was already implemented as complete workflow within nipype (nipype.workflows.dmri.fsl.artifacts.remove_bias), ready to be integrated into the diffusion pipeline.

Diffusion Feature Maps. Depending on the tissue composition in the brain, the diffusion of water molecules is more or less restricted. While diffusion is equal in all direction in cavities such as ventricles, it is more directional in structured tissue like WM fibres. The reason for this are cell structures that form a barrier for water molecules, forcing the diffusion to be more anisotropic in organised tissue compartments. Very briefly, diffusion can be measured by applying gradients of magnetic fields in different directions. To describe anisotropic diffusion in a three dimensional space adequately, at least six diffusion weighted and one non-diffusion weighted image will need to be acquired. From this the proportion of diffusion along the principal axis (diffusion coefficients) can be estimated and described as a symmetric 3×3 matrix called the diffusion tensor. The diffusion tensor can also be represented as an ellipsoid, characterised by the three orthogonal eigenvectors and their eigenvalues λ_1, λ_2 and λ_3. These define the orientation and shape of the ellipsoid. Therefore, diffusion can also be characterised by the eigenvalues' relationship. To express this as a scalar value, different options have been suggested [197]. For example, *fractional anisotropy* (FA) is measuring the anisotropic portion at each position (i.e. every voxel), with high values indication strong directional diffusion (for example in WM).

$$FA = \sqrt{\frac{1}{2}\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}{{\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2}} \tag{2.3}$$

The average of the eigenvalues characterises the *mean diffusivity* (MD), with high values representing less restricted diffusion (for example CSF in ventricles).:

$$MD = \frac{1}{3}(\lambda_1 + \lambda_2 + \lambda_3) \tag{2.4}$$

Although MD is rotational invariant, it can be affected by the choice of b-values [194]. For the pipeline, both were computed via least weighted squares optimisation through nipype's FSL interface for `dtifit`). In addition DWI trace maps[11] were calculated, as they can provide a good contrast for manual lesion delineation:

$$DWI_{trace} = S_0 e^{-b\,MD} \tag{2.5}$$

With b representing different b-values, so consequently, multi-shell diffusion data would have several DWI trace maps. Furthermore, the *anisotropic power* (AP) map was computed. The aim of the AP map was to introduce a noise robust measurement to characterise anisotropy. It has been reported to have a similar contrast to a T1w image, which is why it could facilitate coregistration of diffusion to anatomical scans [47]. This map was defined as the sum of the angular power spectrum of each *spherical harmonic* (SH) of even order l:

$$AP = \sum_{l=2,4,6} \frac{1}{2l+1} \sum_{m=-l}^{l} \|C_{l,m}\| \tag{2.6}$$

[11] not to be confused with *Trace*, which is the sum of the eigenvalues

where $C_{l,m}$ is the coefficient for the SH base of order l and degree m (further explanation to SH in Section 5.1.2). This was computed via the nipype interface (dipy.anisotropic_power).

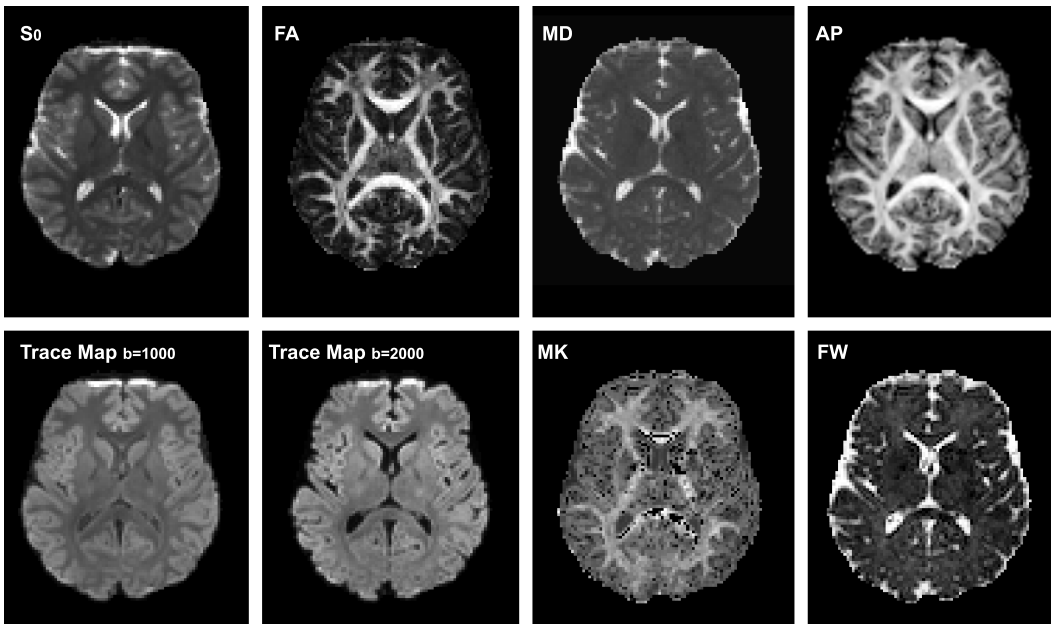


Figure 2.10: Overview of Diffusion Parameter Maps for an Individual Subject

Although standard DWI assumes a Gaussian distribution of water molecule diffusion, this does not hold true for complex biological tissue. This deviation from a normal distribution, becomes stronger when stronger diffusion gradient are applied (i.e. higher b-values) and can be characterised by the kurtosis [113]. The pipeline automatically recognises if any b-values larger than 1500 s/mm^2 are present and in that case models the *diffusion kurtosis imaging* (DKI) tensor to compute diffusion parameter maps such as *mean kurtosis* (MK, see Figure 2.10). Resolving the image of the brain into small discrete units during acquisition can result in partial volume effects, where different tissue types are influencing the signal within one voxel. This is particularly problematic when *free water* (FW) reduces the apparent diffusion coefficients of tissue structures. Different approaches have been proposed to estimate FW diffusion on DWI images. The method integrated in the pipeline does not require any local spatial constraints and works reliably for multi-shell DWI data [100]. So if more than one non-zero b-value exists, the pipeline automatically triggers a custom interface applying Dipy's (version 0.15.0) FW elimination model. This allows to estimate the FW volume (Figure 2.10) and corrected FA and MD parameter maps accordingly (not shown).

Fibre Tract Segmentation. Analogous to the anatomical pipeline, the brain was parcel-

lated based on the diffusion data. For this, the open source tool TractSeg[12] was used as it has been shown to rapidly and accurately segment WM tracts in healthy subjects [263]. The technique is based on a neural network that predicts the labels of 72 fibre tracts. The network was trained on semi-automatically segmented fibre tracts for 105 selected HCP subjects. For this, tracts were first automatically identified after whole brain tractography and refined where needed. After several quality checks and manual adjustments, binary maps for all tracts for all subjects could be generated. Those served as reference segmentations to train a 2-dimensional encoder-decoder CNN. TractSeg has been shown to outperformed other fibre tract segmentation methods (e.g. ReconBunlde, TRACULA or multi-atlas approaches) and is fast to apply. The available pretrained model can be applied via command line, hence was easily integrated as a custom nipype interface.
From the automated segmentations the tract volumes, mean and standard deviation for FA and MD of each ROI were automatically computed (IDP #1). Since the tool provides probabilistic estimates for voxels belonging to a region, both statistical metrics were also calculated with a weighted voxel input. All computed metrics were stored in one CSV file for easy access. Eventually, those single files could simply be concatenated for all processed data within one study.

Coregistration. In order to understand which modality is most beneficial for the coregistration of diffusion parameter maps to T1w image, a short experiment was conducted. First the diffusion MR images from the *Scan-Rescan* database were corrected for artefacts (noise, Gibbs ringing, removal, head motion, and bias-field) and diffusion parameter maps were extracted. The T1w image was bias field corrected and masked (as described in the structural pipeline). Then single diffusion maps (b_0, FA or AP) or different combinations (FA+b_0, FA+AP, FA+AP+b_0)[13] were rigidly coregistered to the masked T1w image (antsRegistration). For this MI within the brain mask was chosen as cost function. The transforms found were then equally applied to the FA map to project it to T1w image space (WelchWindowedSinc interpolation). Thus, the only variable factor was the transforms used. Then, NCC between all projected FA maps and corresponding T1w images were calculated to estimate the registration quality. The impact of using different diffusion parameter maps was very subtle, with average $NCC \approx [0.634 - 0.635]$. Registration with b_0 maps only led to one out of 24 failed coregistration ($NCC = 0.245$), however, the overall performance was slightly worse than all other methods ($NCC = 0.614$). Coregistration seemed to be slightly hampered by inclusion of the b_0 volume (b_0, FA+b_0, FA+AP+b_0), but almost identical for

[12] `https://github.com/MIC-DKFZ/TractSeg`
[13] b_0+AP was not tested

all other combinations. There was a tendency of AP maps alone performing marginal better than FA maps alone or FA+AP maps. Despite that, a coregistration based on FA+AP maps was integrated in the pipeline to leverage the robustness of a multi-parametric approach. The b_0 volumes were not included as they seemed to have an adverse effect on registration quality and including of more image contrasts prolongs the registration process. Besides this, coregistration mostly benefited from T1w images to be masked and starting the alignment with a good initialisation. ANTS allows for an image intensity based initialisation, which led to a robust overall performance. A statistical significant difference between all approaches could not be found (repeated measurement ANOVA: $p = 0.3034$, individual t-tests for repeated samples: all $p > 0.2$). In the end, a multi-modal coregistration of FA and AP maps to T1w images was integrated into the pipeline as it provided qualitatively best results. All diffusion maps (FA, AP, MD, trace maps) were eventually projected to T1w image space.

Spatial Normalisation. Analogous to anatomical scans, spatial normalisation of diffusion images would be beneficial to analyse DTI data via VBM or *tract-based spatial statistics* (TBSS) [232]. Since, tensor-based registration [281] has been shown to outperform intensity-driven registration of diffusion parameter maps [260], the former was favoured to be integrated into the pipeline. However, diffusion tensor registration can be error prone and is most successful when using a study-specific template. Which is why, artefact corrected DWI images are converted to a tensor-format readable by the *DTI-ToolKit* (DTI-TK).[14] This allows to conveniently compute a study-specific template right after all scans of interest have been pre-processed. Such a template is usually not very generalisable and has to be computed for each study individually, which it why it was not integrated in the pre-processing pipeline.

2.4.2 Quality Control Metrics

QC #1: Denosing. To evaluate the noise level of the diffusion scans the SNR was computed. This was defined as the ratio of the mean μ of the denoised b_0 signal and the standard deviation σ of the noise estimated by the denoising algorithm (difference between original and denoised signal). For this, only voxels x within the brain mask M were considered:

$$SNR = \frac{\mu(X_{b_0})}{\sigma(X_{noise})} \quad with \quad X = \{x_1, ..., x_n \mid x \, \epsilon \, M\} \tag{2.7}$$

Examining the metric on all CENTER-TBI DWI scan, the average SNR was found to be lower for controls (N=271, mean [IQR] = 116.4 [97.3]) than for patients (N=2355, mean

[14] `http://dti-tk.sourceforge.net/pmwiki/pmwiki.php`

[IQR] = 153.1 [94.3]). Usually, better signal would be expected for control subjects, however, hyper-intense pathology in patients can skew SNR to higher levels. Generally, a varying SNR was observed across different centres (Figure 2.11), which will need to be taken into account for a multi-centre analysis. While some centres (A, I, K, N) showed an SNR well below the overall average across all centres, some other had a much higher SNR (e.g L). An image with one of the lowest SNR (16.3, centre I, Siemens scanner) had a high variation in noise (σ=17.0) but low b_0 signal (Figure 2.11 A). In contrast, a high SNR scan (1008.7, centre L, Philips scanner) had a low noise variance (σ=5.9), but a twice as high b_0 signal, which was likely a consequence of hyper-intense pathology being present (Figure 2.11 B). The discrepancy between healthy and lesioned brains will need to be taken into account when curating a TBI database.

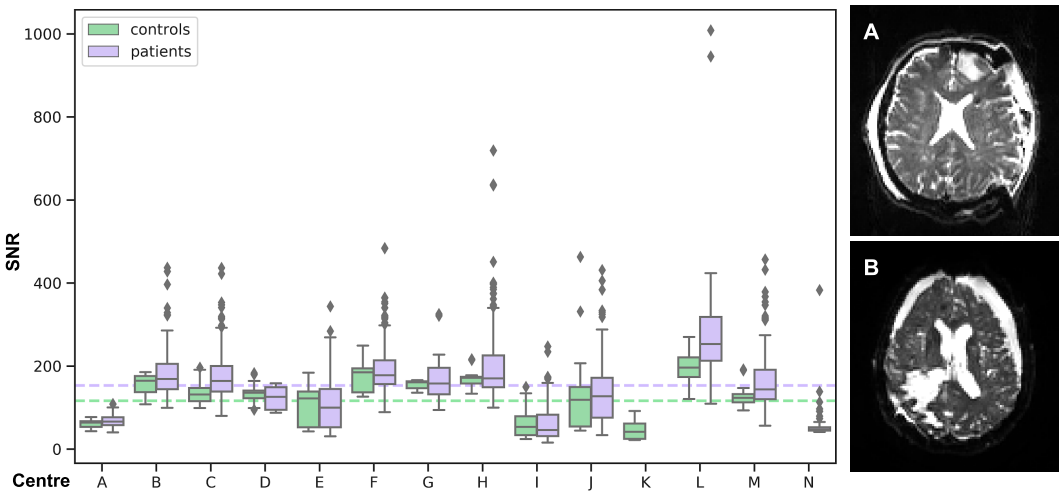


Figure 2.11: Distribution of SNR across Centres. **Left:** The SNR distribution was centre specific. The dotted lines show the average over all controls (green) and patients (purple) **Right:** Images with very low (16.3, A) and high (1008.7, B).

QC #2: Physically Implausible Signal. Diffusion leads to signal decay, which is why for each voxel the signal should be higher in the non-diffusion weighted image than in the diffusion sensitised image. However, because of acquisition artefacts (e.g. Gibbs ringing) this might be violated, leading to measurements of *physically implausible signal* (PIS). With PIS defined as the voxels x within brain mask M for which the signal in the first DWI volume is higher than that in the b_0 volume

$$PIS = \{x \, \epsilon \, M \mid (x_{DWI} - x_{b_0}) > 0\} \tag{2.8}$$

the number of voxels with PIS were counted before (PIS_{before}, that is right after denosing) and after (PIS_{after}) Gibbs ringing removal. To combine both metrics into one, the ratio

of both counts was computed (PIS_{after}/PIS_{before}). A lower ratio would indicate a signal improvement after Gibbs artefact correction. The PIS count ratio was below 100% for all scans, but for the two images with excessive noise corruption (see example later in Figure 2.12 A). Excluding those two scans, the number of PIS voxels after the corrections was slightly higher for patients (mean [IQR] = 945 [794]) than for controls (mean [IQR] = 861 [643]). The PIS ratios for controls (mean [IQR] = 43.7% [16.1%]) and patients (mean[IQR] = 45.6% [21.6%]) were similar and indicated that on average more than half of the voxels with PIS were eliminated by Gibbs correction.

QC #3: Brain Masking. Similar to the quality control for the T1w image brain extraction, two masks are computed for DWI scan as described earlier. The ratio of the MRtrix3 and ANTS brain mask could yield information about the quality of the brain mask. Both controls (mean [IQR] = 98.4% [3.2%]) and patients showed a very similar average ratio (mean [IQR] = 98.9% [3.7%]), however, patients included more outlier scans on both sides of the spectrum. The lowest ratio for controls (75.4%) was the result of MRtrix3 strongly undersegmenting the brain, the ANTS brain mask used for all subsequent processing steps, however, was intact. The 2^{nd} lowest ratio for controls (90.0%) was not noticeably corrupted. The two control scans with the highest ratio (117.6% and 122.4%) belonged to the same subject and showed severe noise artefacts (Figure 2.12 A). Brain masking via coregistration to T1w images seemed to be successful, but sometimes failed when applying MRtrix3. This discrepancy is reflected in the ratio of the brain mask volumes. Other outliers among the controls showed an over- or undersegmentation of MRtrix3. Ratios varied more for patients, and various causes for low QC rations were observed. These included strong brain distortions (Figure 2.12 B: 50.3%), flawed mask back projection from T1w space (Figure 2.12 C: 81.7%) or deformed brains due to injury (Figure 2.12 D: 82.1%, E: 88.8%). Scans with ratio higher than 90% seemed to have only minor discrepancies between both brain extractions. However, scans with high ratios started to show different artefacts, such as signal fading where MRtrix3 grossly oversegmented the brain (Figure 2.12 F: 129.2%). Besides this, a systematic corruption of scans through failed distortion correction could be identified. Although displaying susceptibility artefacts the brain mask (ANTS) could still be computed fairly accurately before the correction (Figure 2.12 G1). However, the failed susceptibility correction left the brain deformed (Figure 2.12 G2) and resulted in a much larger MRtrix3 brain mask. Consequently, this was reflected in a high QC ratio (125.6%). Upon closer visual inspection, a corrupted and misaligned extra b_0 volume (with opposite phase encoding direction) was found. A QC mask ratio within the range 90-110% seemed to indicate sufficiently accurate brain masking.

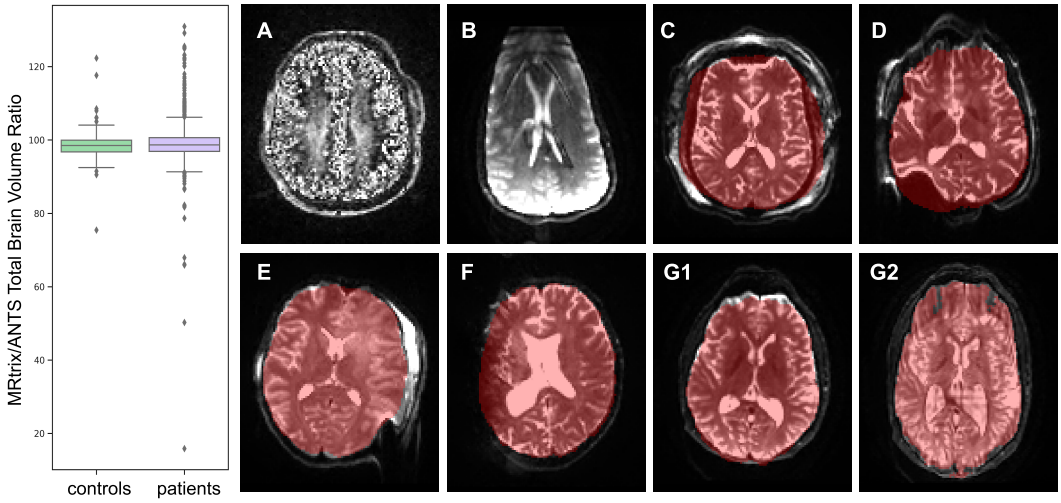


Figure 2.12: Quality Controls for DWI Brain Masking. **Left:** Comparison of controls and patients showed a fairly similar distribution of the QC mask ratio, although, patients showed unsurprisingly more outliers. **Right:** Displayed are different examples of cases with low or high QC ratio and the respective ANTS brain mask (red). Noticeably different ratios could point out flawed image acquisition such as excessive noise (A, 117.6%) or distortions (B, 50.3%), failed brain extraction (C, 81.7%), deformed brains due to brain injury (D, 82.1% and E, 88.8%) as well as scans for which correction of susceptibility distortion failed (G1 and G2, 125.6%). G2 shows the MRtrix3 brain mask (red).

QC #4: Head Motion Parameters. FSL eddy provides as output the estimated average and maximum head motion during the DWI acquisition. This includes both the *total motion*, which is relative to the very first acquired volume (usually b_0), and the *relative motion*, which is the motion with respect to the preceding volume.

Table 2.4: Quality Metrics for Head Motion During DTI Acquisition. Quality control metrics displayed as mean [IQR]

	Total Motion		Relative Motion	
	average	max	average	max
controls	0.364 [0.217]	0.675 [0.361]	0.177 [0.074]	0.514 [0.281]
patients	0.476 [0.264]	0.863 [0.496]	0.193 [0.106]	0.608 [0.340]
t-test	$p < 0.001$	$p < 0.001$	$p = 0.084$	$p = 0.027$

While total motion could help to identify a head drift throughout the scan, relative motion could highlight a subject's sudden head movements. Since head motion can have an impact on diffusion parameters, it is important to detect outlying subjects and a possible bias between control and patient groups. Table 2.4 summarises the QC results for the CENTER-TBI database. Apart from relative average motion, patients were observed to

have a significantly higher total and relative motion compared to controls.

When inspecting subjects with increased QC metrics, obvious head motion was observed. For example, a patient that tilted the head mid-scan, such that all subsequent scans were strongly rotated relative to the first volume, resulting in the highest average total motion (5.082, Figure 2.13 A). Severe head movements also led to inter-scaling signal across slices and striping artefacts (Figure 2.13 B, maximum total motion = 9.183). Such scans should be excluded from analysis. Once corrupted scans are excluded, the motion parameters of the remaining subjects would need to be checked across groups of interest to rule out any group-effects. Eventually, motion metrics may be incorporated as covariates in the analysis.

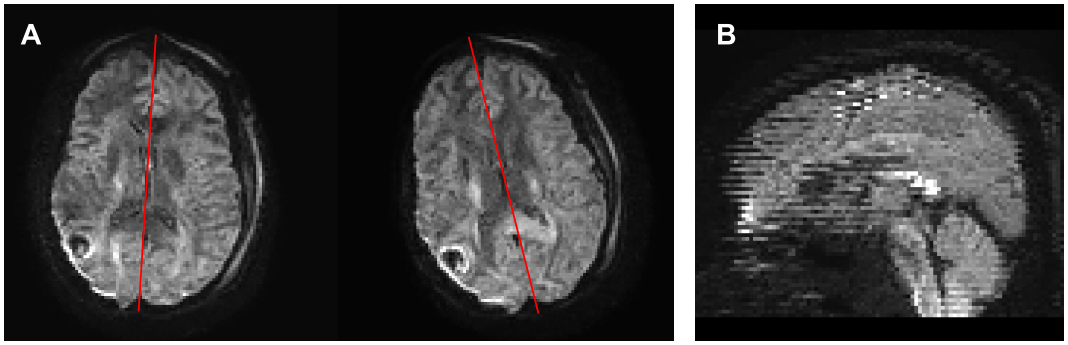


Figure 2.13: Quality Control for DWI Head Motion for CENTER-TBI Subjects. **Left:** Inspecting the patient with the highest average total motion (5.082) revealed a strong tilt of the head mid-scan that was kept through the rest of the scan (A, left: volume #18, right: volume #28, brain axis visualised in red). **Left:** Scans with high motion QC metrics associated also showed striping artefacts, such as the ones visible on the sagittal DWI image slice (B, maximum total motion = 9.183).

QC #5: Coregistration. For joint analysis of T1w and diffusion MRI scans (e.g. FA metrics within MALP-EM ROIs), it is useful to align both to create a spatial correspondence. To monitor the quality of coregistration of FA maps to T1w space, again the NCC between the two scans was computed. Since two different contrast images were compared, a generally lower NCC was expected than for the T1w images spatially normalised to the Cam-CAN template (see Section 2.3.2 QC) #3). For the CENTER-TBI database the average NCC was slightly higher for controls (mean [IQR] = 0.613 [0.115]) than for patients (mean [IQR] = 0.597 [0.119]). The control scans with the lowest NCC could again identify the noise corrupted images as shown before (Figure 2.12 A). Lower NCC cases were found for patients and some of those cases are shown in Figure 2.14. Overall, the coregistration between T1w images and DTI scans was successful and lower NCC values mostly flagged images with pathology. This could be useful to pre-sort scans for visual inspection, but also to exclude those scans from analysis based on processing steps that are likely to fail due to

the presence of lesions (for example brain parcellation). Possibly the NCC could be linked to size and type of pathology.

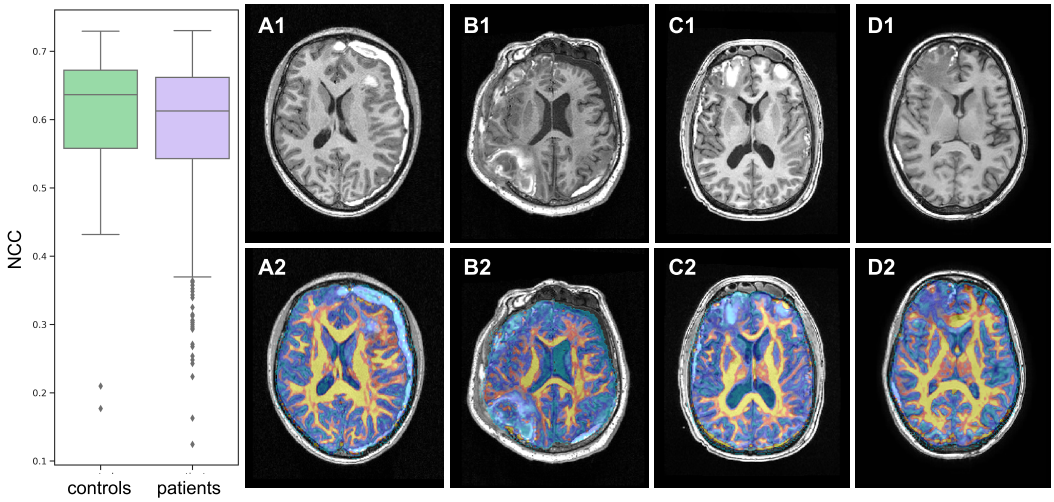


Figure 2.14: Assessment of DTI Coregistration. **Left:** Similar distribution of QC ratio for patients and controls, although patients displayed more outlier cases. **Right:** Top column shows the T1w image (A1-D1) and bottom column shows the aligned FA maps (in colour) overlaying on top of the corresponding T1w images (A2-D2). Cases with low NCC included different types of pathology, such as hyper-intense subdural haemorrhage (A1, NCC = 0.243), contusions in occipital (B1, 0.303) and frontal (C1, NCC = 0.357) lobes and lowered intensity in pathological frontal lobe (D1, NCC = 0.446).

2.5 Discussion

2.5.1 Database Management and Quality Control

As neuroimaging studies grow in sample size, so do the challenges to process and analyse databases adequately. In the past, suitable subjects were pre-selected, based on the clinical questions that should be answered, before processing any of the MRI data. This, however, easily leads to inefficient data handling, because different researchers may choose overlapping subsets of subjects from a larger study. While this not only takes up unnecessary disk space to store duplicates of the database, it is also computationally ineffective to recompute common processes that have already been completed. Moreover, different and sometimes sub-optimal tools would be chosen, as there is a bias towards familiar methods, rather than potentially more complicated state-of-the-art techniques. Consequently, findings from different researchers (even in the same lab) may be less comparable when using different processing tools. Thus, centralising imaging databases should be strongly advocated. Instead of processing a subset of the imaging study, the whole database should be sorted, curated

and processed in the same way and made easily accessible for collaborating researchers. Processing and IDP extraction could change flexibly with new requirements for analysis. Besides centralising data processing, the emphasis lies also on methods for automated quality control. The more data are available the less feasible visual quality assessment becomes. Imagine 1000 scans that shall be processed: Visually checking each input is already tedious, but checking in addition the output of one single processing step already doubles the work load. With many, inter-connected processing steps the complexity increases exponentially, and an initially daunting task becomes intractable.

The quality control metrics introduced in this chapter have been shown to be useful to detect corrupted imaging data or flawed results of processing steps. However, one current limitation is that they are unspecific, as they act more as a surrogate then actually explaining the source of failure. The ratio of the two different brain masks, for example, indicates whether either of the two algorithm failed to extract the brain. However, the reason for failed brain extraction is currently not characterised automatically. Only a visual check can reveal whether the image is completely distorted, includes minor acquisition artefacts or the brain anatomy is too uncommon (e.g. deformed brain due to TBI, or large amount of tissue surrounding the head). In order to improve the pipeline, it is important to assess if a processing tool failed because of a flawed image or an unsuitable algorithm. If an algorithm's performance is found to be systematically inadequate for a challenging dataset, it should be replaced by a more powerful or better tailored tool. On the other hand, if images are corrupted during data collection, they will need to be detected early on to minimise computational costs and call the researchers attention to exclude this scan from any analysis. Therefore, after processing large databases, MR scans that have been flagged by the current QC metrics, will need to be assessed and categorised. Once patterns of problematic images have been identified, more specialised QC tools could be developed and integrated into the pipeline. Different machine learning models have already been suggested for automated quality assessment [17, 160, 63, 145, 235, 87]. Future work will focus on tools that target specific artefacts to first detect whether the images are usable for any further processing. Eventually, each processing step within the pipeline should be assessed for quality in one or the other way. Currently the pipelines only provide QC metrics summarised in a spreadsheet. Although already useful to support subsequent image analysis, machine learning methods could help to flag flawed data and failed processing steps based on the full set of QC measurements. Ultimately, a comprehensive QC report could be generated that enables imaging researchers to confidently set up their experiments and also could be included in publications. Furthermore, QC metrics that flag severe failure cases should evoke the abortion of the pipeline while informing the user immediately in order to save time and

resources.

Although other open-source toolboxes - such as DTIprep [190] - exist, the pipelines presented here were tailored to the needs of clinicians working on TBI neuroimaging data. One of the main advantages of the in-house pipelines are their flexibility that easily allows to integrate any new tools required. For example, DTIprep indeed provides similar pre-processing modules as the diffusion pipeline, but it does not automatically segment WM tracts via TractSeg, which is a fairly novel method. Furthermore, DTIprep seems to only allow to process DTI data, whereas the diffusion pipeline described here also allows to estimate diffusion kurtosis parameter maps. Another advantageous example is that LPCA was used for denoising diffusion images, whereas DTIprep employs the *linear minimum mean squared error* algorithm. The latter has been shown to be inferior, introducing artefacts at frontal areas due to inaccurate noise estimation [166].

2.5.2 The Pipeline - An Ever Evolving Process

The current state of the pipelines has to be understood as one version of many towards the *perfect* workflow, as it will change over time with new demands. Although the presented pipeline has been developed for and applied to different datasets, new databases might come with different challenges and processing requirements. These may need to be addressed with adjusted parameters or completely new approaches. Since novel algorithms and techniques will constantly be developed, the pipeline will need to integrate such improvements to remain state-of-the-art. Moreover, clinicians and neuroimaging researchers will come up with new ideas for image derived features, which may require additional processing modules. All this together asks for agile development of the pipeline, and only if it is considered an ongoing process will it provide the best possible output in future. Some changes may be only enhancements of the pipeline, such as including new IDPs, others might change the state of the data completely (e.g. new methods for denoising [209]). Consequently, this means that databases would co-exist in different versions, depending on the applied processing pipelines. Centralised data management and processing will allow for versions of the previously processed data to be kept, while providing the most recently updated output.

2.5.3 Future Developments

One obvious improvement will be to update nipype to the most recent version, as this will allow to use newly developed interfaces. For example the diffusion pipeline currently integrates a custom interface that accesses MRtrix3 `mrdegibbs` for Gibbs ringing removal. Newer nipype versions have a readily implemented interface for that, which will be used in future

versions of the pipeline.

Many processing tools and analyses are based on an accurate brain masking, however, this is far from trivial for deformed brains due to lesions and other pathology. For structural T1w images the ANTS brain extraction tool seemed to work best, but this accuracy comes with high computational costs. These are caused by intermediate steps, such as tissue segmentation, which results are not used for later processing steps. Additionally, this brain extraction is based on spatial normalisation which is inherently slow. As previously described (Section 2.3.1) this spatial normalisation seemed to be inferior to the subsequent deformable registration results, hence, it will not find any application later in the pipeline. So, ANTS brain extraction could be replaced with a tool that is much faster, but at least as accurate. Alternatives could for example aim to segment GM, WM and CSF separately and combine tissue segmentations retrospectively. However, this approach likely will fail for brains with large TBI lesions, that skew the intensity profile. One possible improvement could be obtained by integrating a predictive brain extraction model. Many new deep learning methods have been suggested to perform skull stripping [52, 105, 133] including on TBI data [221]. Ideally models for brain extraction would be directly linked to brain tissue segmentation [207]. The advantage of such approaches would be the very quick computing time as well as the ability to incorporate lesion information. Even more challenging is the brain extraction on diffusion MRI. Coregistering the b_0 volume to T1w images and projecting the ANTS brain mask from the T1w scans back to DWI space was fast and rather robust. Nonetheless, by design this involves only a rigid registration, which could introduce errors when susceptibility distortions on DTI scans are too severe. Using a learning approach could easily consider information from T1w images, while still being tailored to the exact brain shape (including distortions) on a DWI scan. So, future developments could be to test new machine learning tools for brain extraction for reliability and integrate the most appropriate one. Furthermore, uncertainty measures from the model could be leveraged to estimate the brain mask accuracy. This could replace the repeated computation of brain masks for QC in both pipelines (ROBEX and MRtrix3 in structural and diffusion pipeline, respectively). On the other hand, the overlap between different brain mask could be estimated via Dice scores, a common metric to compute congruence between segmentations maps, and serve as QC metric.

Once an automated lesion segmentation has been integrated in the pipeline, the information can directly be used for spatial normalisation. This has been shown to be advantageous [5, 27] and ANTS conveniently provides an option for cost function masking during registration. Deep learning tools could also be beneficial to accelerate spatial normalisation through deformable image registration. Promising results for neural network frameworks

have been reported to reach accuracy levels close to state-of-the-art methods [14, 140, 185]. Moreover, neural network based spatial normalisation [14] could also help to drastically accelerate registration approaches based on multi atlases, such as MALP-EM. Alternatively, neural networks could be integrated to directly segment structural images [220]. However, further investigation is needed to understand the impact of such approaches on lesioned brains.

Besides that, deep learning could also help to enhance image quality such as removing motion artefacts [144, 195] or improving resolution [237]. Certainly, deep learning tools have improved performances for many MRI processing tasks (e.g. lesion segmentation) and with new developments the boundaries of data processing will be pushed even further. Nonetheless, one pitfall to avoid for future versions of the pipelines is to replace each conventional tool by a better performing neural network. Instead, tasks should be combined to be solved together. For example, rather than having a neural network predicting the brain mask and another segmenting tissue compartments, it would be more beneficial to have one model that does both simultaneously. Although more challenging, aiming for multi-task approaches will provide more efficient and stronger solutions.

This chapter focused on the the modularity of the pipelines and the integration of efficient as well as accurate image processing tools. However, it did not examine the effects of different methods on clinical research questions (e.g. finding differences between patients and control subjects). While it is generally desirable to use robust and precise algorithms, marginal changes in accuracy may not have a strong impact in the grand picture of the clinical data analysis. The influence of different modules on effect size between patients and controls could be investigated in future.

2.5.4 Integrated Lesion Segmentation

One very important branch of the pipeline will be the integration of lesion analysis. Modern machine learning methods such as CNNs have helped to make great progress in accuracy and precision of lesion detection and segmentation. Despite the convincing performance on controlled training and validation datasets, the applicability in a general setting of a pipeline remains complex and poses different challenges. One of them is to achieve a high accuracy on unseen data. Currently, the best results for lesion detection are achieved with supervised deep learning algorithms. This means, however, that lesions of interest will always have to be annotated first to be able to train a strong model. Multi-centre MRI data might vary enough to break an algorithm that had performed well on a validation dataset. So annotations have to be generated either for most centres, or models have to be included that can deal with the different intensity domains [122].

Commonly, CNNs are trained on a given set of image contrasts (e.g. T1w, T2w, FLAIR etc.)[15], which for most network designs means the model can only detect lesions when the exact same contrasts are given (or a surrogate if one assumes two images are interchangeable). For a scan session where one contrast is missing due to inadequate or interrupted acquisition, lesions cannot be detected with such a model in a straightforward way. Considering for example four different MRI contrasts, this leads possibly to 15 different cases where none, one or several scans are missing. Training a model for all scenarios is impracticable and has implications for more complex analyses. That is because lesions are likely to be detected more accurately with four sequences than when some scans are missing. One way to deal with heterogeneous data input for segmentation could be to compute statistical features of the image representation in the latent space of the neural network [94], but this requires further exploration to evaluate its applicability.

2.6 Chapter Summary

Two automated MRI pre-processing pipelines were introduced. The structural pipeline is centred around T1w images, while facilitating the inclusion of other anatomical scans, such as for example T1w or FLAIR images, for flexible application to different databases. The core elements are the brain parcellation via MALP-EM, coregistration of all anatomical sequences to T1w images as well as spatial normalisation to an age-agnostic template created from T1w image of the Cam-CAN study. The diffusion pipeline is based on two pillars: Firstly, the minimisation of image artefacts through denoising and Gibbs ringing removal as well as head motion and eddy current distortion correction. Secondly, the extraction of diffusion parametric maps (e.g FA or MD) and their co-registration to T1w space as well as WM tract parcellation (i.e. TractSeg). Both pipelines extract QC metrics for convenient post-processing data curation. These metrics have been validated on a TBI cohort. Data processed with the pipelines are not only prepared for further analysis, but also extract IDPs that can directly be used for statistical analysis of patient and control cohorts to investigate clinical question.

[15] CNNs can also be trained on a single image contrast if there is only one available

Chapter 3

Application to Mild TBI

3.1 Introduction

3.1.1 Brief Introduction to Traumatic Brain Injury

Risk Factors & Epidemic Character. Traumatic brain injury is defined as an alteration in brain function, or other evidence of brain pathology, caused by an external force [173]. Induced by a sudden event, initial injuries generally could happen to anyone regardless of sex, demographic or ethnic characteristics. However, with traffic accidents, falls, assaults and sport-related concussions being the leading causes for TBI, some groups are at higher risk. For example, men are almost three times more likely to suffer a TBI than women. Another strong risk factor is age, as elderly ($\geq$65) and children (age$<$14) may be more vulnerable to traumatic incidents [75, 161]. It is estimated that approximately 69 million people worldwide suffer a TBI each year [51]. Overall, more TBI incidents were found in high-income countries which, however, may be biased by better reporting systems [51]. With increasing median age, falls becoming proportionally the greatest cause for TBIs in western countries, and a clear change in the epidemiological patterns of TBI has been observed [217]. In contrast, for developing regions - such as Africa or Southeast Asia - road traffic incidents remain the leading cause of TBI closely linked to uprising industrialisation in these countries. The traumatic injury is often only the start of disease progression and recovery from TBI can possibly take years, with some individuals requiring lifelong care and support. In 2016, the annual financial costs associated with TBI were estimated between nine to ten billion US dollars [75]. Due to its ubiquitous character and the wide spanning consequences, TBI has often been described as a *silent epidemic* which creates an immense socio-economic burden globally [161, 217].

Pathological Patterns. Depending on the mechanism and the strength of the initial impact, brain injuries occur with different severities. Although these cover a wide spectrum, patients are usually trichotomised as mild, moderate or severe TBI (approximately 80%, 10% and 10% of all cases, respectively), usually based on the *Glasgow Coma Scale* (GCS) [240].[1] Besides this classification, patients can be grouped according to radiological evidence of brain damage. Primary traumatic injuries can include focal injuries such as cortical contusions, intraparenchymal haemorrhage and subdural or epidural haematomas. Each of these are a consequence of impact site and severity as well as different underlying pathological mechanism. Cortical contusions, for example, mostly appear where the brain makes contact with irregular, prominent bone structures on the inner surface of the skull, and commonly occur in the inferior frontal or temporal lobes. Contusions can be observed both directly at and opposite to the site of impact termed coupe and contre-coupe injuries, respectively. Moreover, vascular damage can result in haemorrhage and haematoma. Besides this, impact forces can cause skull fractures that further damage the brain and can cause for example injury to nerves and arteries (at the base of the skull). Alongside those patterns, diffuse vascular and diffuse axonal injury may occur. These are usually caused by rapid accelerative and decelerative forces. After the trauma and the initial presentation of pathological patterns, a complex cascade of mechanisms causes secondary injuries to the brain. Vascular damage can lead to a formation of new haemorrhages within hours after the TBI, resulting in increased intracranial pressure or hypoxic ischaemia [75]. Additionally, secondary injuries are triggered on a cellular level through an immune system response [111, 229] and on a molecular level as an unregulated flux of ions due to compromised cell membrane integrity [196]. Furthermore, neuroinflammatory processes can take place minutes to months after the traumatic event and the release of neurotransmitters may create a toxic environment exacerbating the initial tissue damage [75, 141]. This heterogeneity in pathological patterns also brings a great variation of symptoms in TBI patients.

Consequences & Outcome. While TBIs can be very different, all grades of severity share common linkages to acute and long-term neuropathologic damage and brain dysfunction [172]. Some symptoms may resolve within days, weeks or months post-injury. However, others such as fatigue, poor cognitive performance, depression or chronic pain may develop and persist for years, drastically impairing life quality. A meta-analysis [224] of 39 studies revealed that cognitive functioning (e.g. memory or attention capacity) is restored most quickly within the first few weeks after mild injury and reaches a plateau during one to three months post-injury. Moderate to severe TBI patients may experience continuing im-

[1] other metrics can be the level of consciousness or post-traumatic amnesia

provement within the first two years after the incident, but their cognitive functions might remain impaired beyond that time point [224, 262]. Although difficult to distinguish from somatic symptoms, depression and social impairment have been repeatedly linked to TBI. Both have been described as consequence after TBI since the mid 1980's [128], and were shown to be present at least up to 12 months post-injury [84]. Symptoms of depression were reported by patients of any injury severity, and may even be pronounced in patients who had suffered a *mild TBI* (mTBI) [204]. Further long-term consequences can be chronic pain and closely linked sleep disorders [265], or other secondary psychiatric sequelae such as post-traumatic stress syndrome [61, 146] or sexual dysfunctions [216]. A long term study showed that, despite high levels of independent daily living, approximately 40% of patients were dependent on more support than before the injury. Moreover, problems that were present at two years after the TBI (e.g. fatigue or balance problem), often persisted until the ten year mark post-injury [202]. Further down the line, TBI have been associated with increasing the risk for neurodegenerative diseases (e.g. Alzheimer's disease) [208, 230]. Finding early biomakers that help to characterise injury better and quantify the likelihood of recovery or disability is crucial to improving patient management, outcome and quality of life.

3.1.2 Role of Neuroimaging for Mild Traumatic Brain Injury

The majority of TBIs (75-85% in 2003) are considered to be mild with a GCS score of 13 to 15 [68]. However, the definition of mTBI has been challenging due to the lack of reliable biomarkers of injury [22] and the influence of various factors (e.g. education and training of the physician observing the patient) on the GCS score [214]. Most mTBI patients experience a full recovery from neurological deficits, but a substantial subgroup (5–30%) develops prolonged neurocognitive problems [45, 172]. Distinguishing those patients with persistent symptoms from those making a full recovery in a mTBI cohort is difficult based on clinical assessment exclusively. Therefore, radiological metrics bear a great potential to characterise injury severity and predict outcome. Nonetheless, while skull fracture and large lesions often mark severe cases, mTBI can be much more subtle. Sudden acceleration or deceleration forces acting on the brain induce deformation and shearing within the brain causing microscopic, multi-focal or diffuse tissue damage on a cellular level. Particularly affected are axons due to their elongated cell structure. Besides cortical contusions and hyper-intensities or micro-haemorrhages linked to shearing related WM injury, most mTBIs appear normal on conventional MR images [188]. Since diffusion MRI is able to characterise the state of cerebral WM integrity [3], DTI has seen a tremendous interest in the field

of TBI research.[2] Substantial evidence of the effectiveness of DTI in detecting TBI has emerged over the last decade. This established the consensus of lowered FA in WM as a characteristic of TBI-related abnormalities [102]. Decreased FA can reflect less isotropic diffusion as it can be observed in traumatic axonal injuries [88], which may potentially indicate the disruption of the integrity of WM fibre tracts. For example lower FA values within some ROIs (i.e. internal and external capsules, or *corpus callosum* [CC]) were found to correlate with worse three and six months outcome after injury [276]. A systematic review revealed that an unfavourable outcome in TBI with diffuse axonal injury was three times more likely compared to TBI in the absence of such axonal pathology [250]. Another meta-analysis of mTBI studies, confirmed the decrease or increase of FA and MD, respectively, within the CC [9]. Further examinations of mTBI using DTI have shown that frontal and temporal WM pathways are predominantly affected. Such micro-structural changes in WM could be linked to behavioural and cognitive performance measurements [188, 157]. In fact, severities of traumatic axonal injury and TBI seem to correlate. Physical alterations found in different regions such as the CC, fornix, subcortical WM, as well as the cerebellum are believed to contribute to the extent of symptoms after mTBI [172]. Although aberrations of DTI metrics could be associated with specific brain locations and patient outcome, the heterogeneity of patient cohorts have hampered the generalisability of abnormal DTI findings [102]. A recent meta-analysis summarising 42 studies reaffirmed previous discoveries of lower FA and higher MD values in TBI subjects and reported more subtle WM changes for mild patients in comparison to moderate and severe TBI cases. Moreover, the measured differences in DTI parametric maps seemed widespread[3], indicating global WM damage after TBI [259]. Imaging derived diffusion characteristics were repeatedly reported to be useful biomarkers for WM injury [211] and differentiate TBI at a group level. However, there is still a lack of strong evidence to prognosticate patient outcome at the level of individual patients [58, 261]. One step in this direction could be the analysis of multi-modal images, including volumetric measurements from structural MR (i.e. T1w scans) and quantification of abnormal diffusion within the brain [246]. Besides cross-sectional analysis, longitudinal studies could provide further information for disease progression in TBI patients.

3.1.3 Related Work for Analysis of Mild TBI MRI Data

Although not exhaustive, this section aims to provide a broader overview of volume and diffusion changes in the brain after mTBI.

[2] Number of *Google Scholar* results for "DTI mTBI" (Jan. 2020): 1480 in 2001-2010; 7720 in 2011-2020

[3] abnormalities were observed for 35 different ROIs across the studies included for the meta-analysis

Volumes. Structural brain volume changes in mTBI have been investigated in several reports, and overall highlighted the influence of time after the traumatic incident on development of tissue atrophy. A comparison of mild and moderate TBI patients three months after injury found no differences in brain parenchyma or CSF volume when compared to healthy controls [163]. However, global brain volume loss was detectable in mTBI patients one year after the injury, with atrophy observed both in GM and WM [284]. Moreover, a decrease in brain volume was found to be present in the chronic stage (approximately two years post-injury), with regional volume losses in the forebrain, cerebral WM and cerebellum appeared to be affected [218]. More recently, significantly reduced cortical and subcortical volumes (in the nucleus accumbens, and caudate) were found in mTBI when comparing one year follow-up scans to the initial image obtained one month post-TBI. [98]. These findings may suggest that tissue loss in mTBI evolves over time. Nonetheless, a longitudinal study has reported regional volume decrease (in the caudate, putamen or thalamus) as early as two month after injury [278]. And another study found that mTBI patients experience significant cortical volume loss and an increase in ventricle size within one month post-injury [247]. Furthermore, mTBI patients were found to have a significantly higher atrophy rate than controls a few months (>3) after the injury [163]. While discrepancies have been repeatedly found between patients and controls, differences between patient groups with varying symptoms or outcome is less consistent. For example, there were no significant longitudinal changes observed when comparing patients with *complicated* or *uncomplicated* mTBI. [98]. However, volume of brain parenchyma was observed to decrease more in patients with loss of consciousness [163]. In addition, an elevated rate of tissue atrophy was linked to the inability to return to work [218] and decreased ROI volumes were associated with neurocognitive performance [284].

While brain volume loss in mTBI patients has been observed multiple times, increases in the volume of brain regions has also been reported, but far less frequently. A very recent study has found abnormally enlarged cortical GM areas in mild to moderate TBI patients when compared to a large database of control subjects. At the same time, however, the authors found tissue atrophy in cerebral WM [219].

General volume loss in mTBI patients has been well established, but most of the above mentioned studies included less than 20 [163, 218, 247, 284] or up to approximately 60 [98, 219, 278] mild and moderate TBI patients. A small sample size introduces biases associated with patient selection and reduces the robustness of these studies and limit their generalisability. In fact, a recent review [93] concluded that observations of brain tissue atrophy after mTBI remain inconclusive, but moderate and severe TBI patients show more distinct patterns of general and focal atrophy on a global or regional level. More specifi-

cally, this includes widening of cortical sulci, cortex thinning, shrinking of the hippocampus and increase in ventricle size [93]. The underlying mechanisms for brain atrophy are direct disruption of structural cell components, but also a cascade of secondary responses on a molecular level [93]. While the incidence of a TBI may have a direct impact on brain tissue loss, it also seems to trigger an ongoing process of accelerated tissue atrophy months and years after the insult [20, 39].

Diffusion. Besides volumetric changes in anatomical brain regions, mTBI has frequently been associated with abnormal water diffusion within the brain. This reflects the disintegrity of axons in the WM, which are particularly vulnerable to mechanical loading of brain tissue during rapid head accelerations due to their elongated cell structure and anisotropic arrangement [116]. The mechanical impact on the WM can lead to disconnection of axons at the time of injury, however disruption is mostly induced by secondary damage caused by axon swelling [116]. Progressive degradation of WM has been shown to extend long after the impact, resulting in loss of brain connectivity in the entire brain [20]. White matter alterations in mTBI can be subtle and highly dependent on the examined cohort and the choice of analysis [103]. However, a comprehensive meta-analysis summarising 28 DTI studies showed the robust findings of significant FA reduction and increase in MD in the CC after mTBI [9]. The splenium was mostly affected, with marginal and no significant differences for the mid-body or genu of the CC, respectively [9]. Nonetheless, measuring differences in WM integrity in TBI is far from trivial as diffusion metrics change dynamically with *days post-injury* (DPI) [227].

Early on, DTI metrics were analysed for mTBI patients at different time points. Inglese et al. [107] have examined two patient groups, scanned in the acute or chronic phase (approximately 4 DPI or 6 years post-injury, respectively). At both time points, patients showed significantly reduced FA and increased MD (splenium, internal capsule) compared to controls. A link to neuropsychological data was not reported for this study [107]. Lowered FA values have also been reported for the *corticospinal tract* (CST), sagittal stratum and *superior longitudinal fascicle* (SLF) in chronic (approximately 7.5 years post-injury) mTBI patients [138]. Kraus et al. [138] could show that the number of WM tracts in TBI patients with lower FA than the control group had worse executive functions and memory. However, for this they have pooled patients scans from both available time-points (approximately 7.5 and 10 post-injury), which does not allow the assessment of progression of neuropsychological functions over time. In contrast to examining patients at a very chronic stage, mTBI has also been investigated on hyper-acute MRI scans (on average 10 hours post-injury). This has revealed a similar pattern of decreased FA and MD within in the splenium, internal

capsule or *cingulate gyrus* (CG) in patients compared to controls [253].

Besides contrasting patients against controls, different patient groups have also been compared to one and another. Patients that showed intra-cranial lesions at admission also displayed decreased FA at follow-up. However, patients without visible lesions, demonstrated no abnormal diffusion in comparison with controls [276]. Despite patients deviating from controls, diffusion was not found to be different for two separate patient groups scanned in the acute (approximately 4.5 DPI) and chronic phase (approximately 6 years post-injury) [107]. Patients dichotomised as achieving good or poor outcome according to their functional status at follow-up (three months post-injury) showed different diffusion patterns on MRI scans approximately two weeks post-trauma [174]. While those patients with good outcome had similar DTI measures to controls, patients with poor outcome showed increased MD compared to controls as well as patients with good outcome. Regions for which such differences were observed included the CC, the *inferior fronto-occipital tracts* (IFOs), the *anterior thalamic radiations* (ATR), the CST as well as the superior and *inferior longitudinal fascicles* (ILFs). Although no differences were discovered between groups when examining FA, this provided evidence that abnormal diffusion measured on sub-acute scans could be useful for TBI outcome prognosis [174]. No changes were detected between sub-acute and 3-months scans [174]. Likewise no variation was found between the two patient groups scanned during the acute or chronic phase [107]. These data suggest that DTI can detect both initial microstructural injury and late tissue status following TBI.

Although many studies replicated the findings of lowered FA and increased MD, discordant results have also been reported. For example, increased FA and reduced MD values were found in smaller mTBI cohorts both during the hyper-acute (approximately 3 DPI) [270], semi-acute (6 DPI) [267] and sub-acute stage (within 21 DPI) [156].

Longitudinal Diffusion Changes. With the growing awareness of time-dependence of TBI, recent studies have addressed changes of diffusion in WM over time. A tract-based comparison of hyper-acute (approximately 24 hours) and three-months scans did not reveal any differences in DTI metrics. With initially elevated MD for the mTBI cohort, this indicated the persistence of WM damage few months post-injury [184]. The comparison of DTI data from two and 12 months after injury has shown different patterns for FA and MD in mTBI patients. While some fibre tracts (e.g. CC, CST) showed decreasing FA and increasing MD over time, others (e.g. internal capsule, ILF and SLF) showed decreased MD values between the two visits [19]. Longitudinal changes both in FA and MD were more pronounced in patients than in the control group [19]. A recent longitudinal study showed initially lower FA (internal capsule, IFO) during the semi-acute phase (seven DPI), which

however, seemingly recovered to baseline one month after injury. Other tracts (e.g. CC) showed evidence of evolving injury, detected by progressive changes up to three months after the TBI [273]. In contrast, in another report, patients with mTBI were found to show initial regional elevations in FA during the semi-acute phase (on average six or 12 DPI), which partially recovered to normal levels within three to five months [171] or a year [43] after injury. Moreover, the examination of a smaller mTBI cohort on several scans within the first eight days revealed complex changes in diffusion patterns with patient-individual trajectories [268].

A substantial amount of research has focused on mTBI, establishing the picture of brain atrophy after injury. Generally the CC seems to be one of the most affected locations of abnormal diffusion in mTBI patients [103]. However, many earlier studies only included around 20 patients [107, 138, 171, 174]. Although more recent studies increased the patient numbers [19, 43, 184, 253, 273, 276], the largest sample size of 80 patients is still very low given the vast heterogeneity of the mTBI patient cohort. In addition, the variation in time points and inconsistency in the chosen analysis tools hamper comparison across studies. A key factor that needs to be taken into consideration for an adequate interpretation of DTI metrics seems to be the time post-injury [48]. Despite the existing efforts, more research is needed to gain a better insight in disease progression after mTBI.

3.1.4 Aims

Although more recent studies have started to recruit larger number of patients, most previous observations are based on small TBI cohorts. Undoubtedly, this limits the generalisation of the findings. To date, there is seemingly only one combined analysis of TBI data from more than one site. This, however, enrolled less than 15 patients from two out of the three centres (104 patients in total). Moreover, this study relied on visual inspection of images, rather than quantitative metrics [228]. Changes in regional volumes (derived from MALP-EM) and DTI metrics (tract count) have been investigated before, but only addressed a smaller cohort of moderate to severe TBI patients [186]. Therefore, the aim for this chapter is to retrospectively build a multi-centre database to examine a larger cohort of mTBI patients. The analysis entails volumetric measures obtained from the MALP-EM atlas as well as regional FA and MD metrics derived from the relatively new TractSeg WM atlas. Besides investigation of site-specific differences, results are presented for three individual databases and their joint analysis. The clinical aim of the study was to assess whether features derived from acute MRI scans are different for controls and patients with good or poor outcome. Furthermore, the prognostic value of regional volumes and diffusion for functional outcome

was examined. Lastly, a longitudinal analysis of all acquired scans looked at the disease progression in patients and whether it differed for patients with good or poor outcome.

3.2 Data Acquisition, Processing & Curation

3.2.1 Databases

Data from three separate TBI studies were examined independently, as well as pooled together for a retrospective multi-centre analysis. The following section provides information for acquisition parameters and numbers of scanned subjects in each of these studies.

Cambridge Database. Mild TBI patients were scanned on a 3T Siemens Verio Magnetom at the Wolfson Brain Imaging Centre in Cambridge, UK. Structural MR images, T1w MPRAGE with FOV: 256×240×192 and isotropic $1\,mm^3$ voxels, were acquired with $TE = 2.98\,ms$, a $TR = 2300\,ms$, and $TI = 900\,ms$ and a 9° flip angle. Diffusion weighted images included 63 different directions on a single shell ($b = 1000\,s/mm^2$) and five non-diffusion sensitised images ($b = 0\,s/mm^2$) evenly distributed across the acquisition time. A 2-fold acceleration factor was applied (GRAPPA) to obtain images with isotropic 2×2×2 mm^3 voxels (63 axial slices; FOV = 96×96) were all acquired with $TR = 11700\,ms$ and $TE = 106\,ms$. Additional data (e.g. b_0 volumes with reversed phase encoding direction) to correct for susceptibility artefacts were not collected. The database included 20 healthy and 22 trauma[4] controls as well as 51 patients. These were scanned at different time points up to approximately 24 months post-injury to capture the TBI progression. With those follow-up scans of the the 51 patients, the databases encompassed a total number 183 patient scan sessions.

Trondheim Database. Structural and diffusion MR images were also collected for 156 mild TBI patients along side 83 healthy control subjects. Subjects underwent imaging on a 3T Siemens Skyra scanner upon admission at the St. Olavs University Hospital Trondheim, Norway, and at follow-up times of three and 12 months. This longitudinal study eventually included 402 and 206 scan sessions for patients and controls, respectively. All those scan sessions included a T1w and DWI scan. The T1w MPRAGE scans were acquired with $TE = 4.21\,ms$, a $TR = 2300\,ms$, and $TI = 996\,ms$ and a 9° flip angle. T1w images entailed 176 slices covering a FOV of 256×256 with isotropic $1\,mm^3$ voxels. Diffusion weighted scans included four baseline volumes ($b = 0\,s/mm^2$) and 60 diffusion sensitised volumes with two

[4]trauma controls were patients who experienced an external trauma that has not directly affected the head, e.g broken leg

different b-values ($b = 1000, 2000\,s/mm^2$, same 30 directions each). Scanner parameters were set at $TR = 8800\,ms$ and $TE = 95\,ms$ to image DWI images (60 axial slices; FOV = 96×96) with isotropic 2.5×2.5×2.5 mm^3 voxels. Additional b_0 volumes with opposite phase encoding directions were acquired (otherwise same parameters).

Turku Database. From the TBI study at the Turku University Hospital (Tyks), Turku, Finland, 114 mTBI patients and 30 trauma controls subjects with both T1w and DWI scans (3T Siemens Verio) were selected based on GCS scores (GCS $\geq$ 13). Follow-up scans around ∼6-8 months post-injury make this a longitudinal database consisting of 204 patient and 51 control scan sessions. The T1w scanning parameters were $TE = 2.98\,ms$, $TR = 2300\,ms$, $TI = 900\,ms$ and a 9° flip angle. The 176 slices covered a FOV of 256×240 isotropic $1\,mm^3$ voxels. The 65 volumes of the DWI scan included one baseline volume ($b = 0\,s/mm^2$) and 64 diffusion sensitised volumes ($b = 1000\,s/mm^2$). Isotropic images (voxel size: 2×2×2 mm^3; 81 axial slices; FOV = 96×96) were acquired with $TR = 11700\,ms$ and $TE = 106\,ms$. Additional scans to correct for susceptibility artefacts were no collected.

3.2.2 Specifications of MRI Processing

All images were processed with both the structural and diffusion pipelines, as described in the previous chapter. The additional b_0 volumes acquired alongside the DWI scans for the Trondheim database were used to correct the diffusion scans for susceptibility distortion. Both other databases did not include such a scan. The Trondheim databases included multi-shell diffusion imaging with higher b-values, which would allow employing more sophisticated models to calculate diffusion in WM microstructure. However, diffusion data from Cambridge and Turku were collected using single-shell acquisition, restricting the diffusion estimation to standard tensor modelling. With the aim in mind to compare and combine all three databases, the same tensor fitting model was applied to all data (FSL dtifit). Since the assumption of linearity of the logarithmic signal attenuation does not hold true for higher b-values [176], the diffusion tensor model was only fitted to the lower ($b = 1000\,s/mm^2$) single shell for Trondheim scans. Moreover, to consider a similar number of non-diffusion weighted images ($b = 0\,s/mm^2$) for the tensor estimation, b_0 volumes after the acquisition of the first shell ($b = 1000\,s/mm^2$) were disregarded. So, only one b_0 volumes was included for Trondheim, despite several b_0 volumes having been acquired. All b_0 volumes have been used for Cambridge data, as they were interleaved between the diffusion sensitised scans.[5] Study data from Trondheim and Cambridge included mTBI subjects only. Turku patients were defined as mild when their GCS score was larger than 12 (GCS $\geq$ 13). Only controls

[5]This was a choice based on practicability, but is acknowledged as a limitation of this study.

and mTBI patients that had both a T1w and a DWI scan were included. The extracted IDPs were regional volume (as provided by MALP-EM) and average diffusion metrics (FA and MD) within TractSeg parcellations. These tools were chosen as MALP-EM has been shown to cope well with TBI-related deformations [151] and ROI based approaches were found to be more sensitive than voxel-wise DTI analysis [103]. Although other structural scans were acquired and processed as part of the automated pipeline, they were not considered for this analysis. Assuming latent axonal injury, patients with visible pathology (in particular lesions visible on T1w scans) were excluded to avoid any bias due to brain parcellation failures.

3.2.3 Data Curation Prior to Analysis

The diffusion processing pipeline provides QC metrics as output (Section 2.4.2) that help to identify problematic scans for which, for example, the acquisition was corrupted. Any scan that was associated with exceeding or unexpected QC measures was visually examined. Cut-off values were empirically chosen for each QC metric individually based on the metrics observed in the whole mTBI database. Once a scan was rejected, its QC metrics were not considered anymore for the subsequent curation. At first, the PIS ratios were checked. Highlighted by a low value (PIS ratio = 66%), one Cambridge scan was excluded as it showed strong striping artefacts due to head motion. Both available scans of one Turku control were picked up on: one showed visible ghosting artefacts outside the brain mask (PIS ratio = 57%), the other displayed noticeable hyper-intensities (PIS ratio = 53%). As the scan quality within the brain seemed unaffected, both scans were kept for analysis. Then, the SNR of the non-diffusion weighted images were analysed. None of the retained scans showed any obvious artefacts. It is noteworthy, that nine scans with the lowest SNR were all collected in Cambridge, while most scans with the highest SNR were from the Trondheim database. Checking the ratio of the actual and control DTI brain mask highlighted a few subjects. Some of those with high ratio were picked out, because the control algorithm (MRtrix3) failed and oversegmented the brains. Since the actual skull stripping algorithm (ANTS) performed as expected, flagged scans were not excluded. Similarly, at the lower bound of the mask ratio, the control mask (MRtrix3) severely unsegmented the brain, but the used mask (ANTS) was acceptable. However, two Trondheim scans highlighted by a low ratio showed strong signal drop-out in the frontal part of the brain. Both scans were rejected to avoid inaccurate results. Two further scans demonstrated very high intensities in the cerebellum, which seemed to have neither directly affected brain masking nor tensor fitting. Therefore, they were kept for the analysis. The scans with the lowest NCC between FA and T1w images after coregistration were checked. Two scans (NCC = 0.38 and 0.43)

were excluded from the analysis due to observable lesions on the DTI scans affecting the WM parcellation. The corresponding follow-up scans were also inspected, and one of them was excluded for the same reason. Remaining scans that were highlighted by greater head motion QC values showed no apparent motion artefacts, such as striping or blurring, upon visual inspection.

The curation process described above led to excluding a total of six scans (one Cambridge, two Trondheim, three Turku). Quality control values were checked for significant differences between the healthy and trauma controls from Cambridge. Since they could not be distinguished statistically ($p > 0.3$ for all t-tests), both categories were considered as a single group for subsequent analysis. Table 3.1 summarises the number of available subjects and scan sessions after the curation process.

Table 3.1: Number of Subjects and Scans of the mTBI Databases

Category	Cambridge	Trondheim	Turku	All
controls (subjects/scans)	42/47	83/206	30/51	155/304
patients (subjects/scans)	50/182	155/400	112/201	317/783

3.3 Experiment Setup

3.3.1 Data Categorisation

Brain tissue atrophy and changes in WM are a consequence of healthy ageing [67]. This is why the age of a subject at scan time needs to be taken into account for any analysis measuring brain volume loss and WM integrity. Dependent on study design and availability, cohorts can vary in age and the time point of imaging after injury across different sites. Plotting the distribution of age at scans revealed that overall Trondheim included younger subjects than both other databases. While Trondheim and Turku showed matching distributions of age for patients and controls, Cambridge patients were spread over a wider range than control subjects (Figure 3.1 left). When combining the databases, the age distribution was mostly driven by the large Trondheim dataset. Pathological patterns in TBI vary over the course of the injury (Section 3.1.3). Therefore, it is important to consider the time point of image acquisition for analysis of TBI patients. All three TBI studies were setup with a different MR collection scheme. In Cambridge patients were scanned at five distinct time points over the course of two years. Trondheim focused on acquiring a hyper-acute scan (DPI$\leq$3) as well as two follow-up scans at three and 12 months. Turku collected scans at mostly the acute phase and six months after the injury, as well as a few follow-up scans one

year post-injury (Figure 3.1 right). Five time frames were defined based on the scan time point distribution and clinical relevance: Acute phase[6] (DPI $\leq$ 42), three months (42 $<$ DPI $\leq$ 150), six months (150 $<$ DPI $\leq$ 300), 12 months (300 $<$ DPI $\leq$ 500), and chronic phase (DPI $>$ 500).

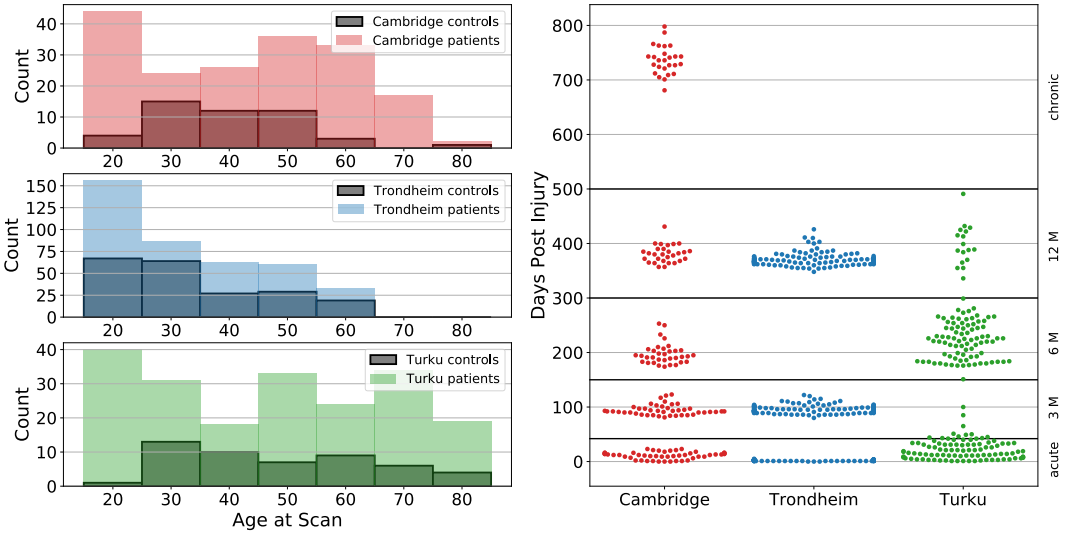


Figure 3.1: Overview of Age and Scan Time Distribution. **Left:** The distribution of ages at scans shows that the Trondheim cohort included younger subjects that the other two TBI databases. Patients and controls scanned at Cambridge showed a different age distribution. **Right:** All three studies collected scans at different time points. Cambridge patients were scanned from the *acute* to *chronic* phase. Trondheim subject were scanned very early on and at the three (*3 M*) and 12 months (*12 M*) mark post-injury. Turku patients were mostly scanned during the acute and 6 months (*6 M*) periods after injury (each point represents a patient scan).

3.3.2 Region Selection for Analysis

Volumetric Analysis. Regional volumes within the brain were estimated via MALP-EM (Section 2.3.1). As summarised previously (Section 3.1.3), mTBI has been associated with regional volume loss at different stages post-injury. This analysis focused on similar anatomical regions, as found in the MALP-EM atlas including: Nucleus accumbens (ROI #3 & #4) [98], caudate nuclei (ROI #8 & #9) [278], cerebellar WM (ROI #12 & #13) [218], lateral ventricles (ROI #23 & #24) [247], putamen (ROI #27 & #28) [278], thalami (ROI #29 & #30) [278], anterior CG s(ROI #41 & #42) [284] and precuneus (ROI #101 & #102) [284]. Besides these ROIs, the cerebral WM [218], cortical GM [247] and ventricles were analysed as a whole [278].

[6]including all three stages of hyper-, semi- and sub-acute

Diffusion Analysis. Analogous to the volumetric analysis, regional diffusion was compared between patient groups and the control group. As described above (Section 3.1.3), differences in diffusion pattern were found in various regions at all phases after the injury. The analysis examined 24 regions that were previously found to be affected by mTBI. These were segmented by TractSeg and included: The ATRs (ROI #2 & ROI #3) [174], and in both hemispheres the cingulum (ROI #12 & ROI #13) [253], the CSTs (ROI #14 & ROI #15) [19, 138, 184], the IFOs (ROI #24 & ROI #25) [174, 273], both ILFs (ROI #26 & ROI #27) [19], the $\text{SLFs}_{I,II,III}$ (ROI #35-40) [138] and the *uncinate fascicles* (UF: ROI #43 & ROI #44) [171]. Furthermore, the CC as a whole (ROI #45) [19, 103, 174, 184] and its sub-parts such as the *rostrum* (ROI #5), *genu* (ROI #6) [142, 171], *anterior* and *posterior mid-body* (AMB: ROI #8 and PMB: ROI #9, respectively) [273] and the *splenium* (ROI #11) [9, 107, 142, 253] were included. Although the internal and external capsules [107, 184, 253, 273] as well as the corona radiata [171, 273] were also reported to show deviating diffusion they were not included, since they are not explicitly segmented by TractSeg.

3.3.3 General Statistical Analysis

After scan selection (see below), volumes or diffusion measurements were harmonised across sites by standardising the distributions for each ROI separately. For this, mean (μ) and standard deviation (σ) of the control population within a centre were computed at first. Then, Z-scores (z) for the particular cohort were derived by subtracting the mean from the metrics (x) and dividing by the standard deviation ($z = (x - \mu)/\sigma$).

Linear regression analysis was chosen to identify differences between subject groups while considering age, sex and acquisition site as confounding factors. Since an initial test showed that a simple linear model with least square fitting resulted in non-normally distributed residuals, a *generalised linear model* (GLM) was employed (statsmodels python library) to fit a Gamma distribution. Non-linear models were not tested to keep statistical model as simple as possible. To respect the domain of the Gamma family of real positive values, all Z-scores were shifted towards positive values with a global constant (i.e. subtracting the global minimum Z-score per ROI metric and adding a vanishing small constant such that the minimum Z-score was just above zero). These Z-scored data were used for all experiments. To account for multiple comparisons, all results were corrected for *false discovery rate* (FDR) via the Benjamini-Hochberg method (statsmodels) and reported as statistically significant when the corrected p-values were below 0.05. False discovery rate was chosen over Bonferroni, as the latter is more conservative and may produce false negatives when correction for a large number of tests. Throughout this chapter, the terms *prediction* and *prognosis* are used to describe the fitting of statistical models to regress imaging data (mostly collected in

the acute phase) against the functional outcome (i.e. GOSE assessed several months after MR acquisition). In other words, no prediction in the sense of data science was performed (e.g. cross validation).

3.3.4 Site-Specific Biases

Since data were included from three centres, at first differences between control subjects were analysed for regional volumes and diffusion metrics (i.e. FA and MD). For this, a GLM was fitted to the control data (subject i) to assess the influence of the acquisition site on the ROI metric (Y_i) while accounting for age at scan and sex. These covariates were chosen as they are know to affect volumes and diffusion in the brain and, furthermore, are widely available.

Model #1: $$\mathrm{f}_\gamma(Y) = \beta_0 + \beta_1 Site_i + \beta_2 Age_i + \beta_3 Sex_i + \varepsilon_i$$

with f_γ representing for the log link function to fit a Gamma distribution, $\beta_{0,\ldots,3}$ representing the parameters to be fitted by the GLM and ε_i is the error unexplained by the model. A separate GLM was employed for each ROI individually.

Besides applying a GLM to all three databases pooled together, all combinations of two different databases were examined via regression analysis as well to identify the impact of different centres. To visualise regional differences, another GLM was fitted analogously, but excluding *Site* as predictive variable. That way, the GLM residuals represented the site-specific differences while adjusted for age and sex. Table 3.2 lists the number of controls included.

3.3.5 Acute MRI Differences between Controls and Patient Groups

The aim was to examine databases in a joint analysis to investigate whether IDPs from the acute phase can be associated with patient outcome. Patients were dichotomised according to their *extended Glasgow Outcome Scale* (GOSE) [269] score into groups with good (GOSE=8) and poor (GOSE<8) functional outcome. Clinically it is more of interest if a patient will show symptoms later on (GOSE<8) or has fully recovered from the TBI (GOSE=8), rather than identifying how bad the symptoms are (GOSE 3-7). Furthermore, combining all patients with unfavourable outcome (GOSE<8) improves the class imbalance. Even when pooling them together, patients with poor outcome only made up 39% of the whole patient cohort (Table 3.2). Examining Cambridge patients with GOSE scores at both

three and six months showed that eight out of 37 subjects had a changing outcome (good vs. poor) between both time points. Seven of those patients changed towards a good outcome, showing a general trend towards improved outcome from three to six months. So, the assumption was that patients functional outcome is mostly stable after three months, but if it was changing it seemed more likely to improve. Thus, GOSE scores from either three and six months were used, since Trondheim and Turku only measured GOSE at three or six months, respectively. In case GOSE scores had been recorded at both time points (i.e. in Cambridge database) the three-months GOSE score was used as it was available more often. In case a patient had multiple scans within the acute phase, the earliest scan was chosen. Eighteen patients had to be excluded, because they had no GOSE score recorded. Any missing records of GCS scores were imputed for patients with the majority value (i.e. GCS=15). Table 3.2 shows the number of subjects included. The employed model included *Category* as predictive variable, indicating a subjects belonging to the control or patient group with either good or poor outcome, while also accounting for age, sex and site as confounding factors.

Model #2: $$\mathrm{f}_{\gamma}(Y) = \beta_0 + \beta_1 Category_i + \beta_2 Age_i + \beta_3 Sex_i + \beta_4 Site_i + \varepsilon_i$$

with f_γ representing for the log link function to fit a Gamma distribution, $\beta_{0,\dots,4}$ representing the parameters to be fitted by the GLM and ε_i is the error unexplained by the model. The continuous dependent variable (Y) was replaced with the volume or diffusion metric of a particular ROI. An individual GLM was fitted for each ROI.
To visualise regional differences, a GLM was fitted analogously, but excluding *Category* as predictive variable. That way, obtained GLM residuals represented the category-specific differences while adjusted for age, sex and site.

3.3.6 Prognostic Value of Acute MRI for Mild TBI Outcome

To test whether regional metrics derived from acute MRI have a prognostic value to differentiate patients with good or poor outcome, a GLM was fitted considering age, sex, acquisition site as well as GCS and DPI as covariates. Control subjects were excluded.

Model #3: $$\mathrm{f}_{\eta}(Y) = \beta_0 + \beta_1 ROI_i + \beta_2 Age_i + \beta_3 Sex_i + \beta_4 Site_i + \beta_5 GCS_i + \beta_6 DPI_i + \varepsilon_i$$

with f_η standing for the logit link function to fit a Binomial distribution, $\beta_{0,\dots,6}$ representing the parameters to be fitted by the GLM and ε_i is the error unexplained by the model. The binary dependent variable (Y) indicated the patients' good or poor outcome.

Table 3.2: Overview of Available Data for MRI Analysis

	Category	Cambridge	Trondheim	Turku	All
Controls	# Subjects	42	83	30	155
	Sex (M/F)	23/19	50/33	14/16	87/68
	Age	39 [18,78]	33 [16,59]	51 [22,91]	38 [16,91]
Acute Patients	# Subjects	39	134	82	255
	Sex (M/F)	28/11	84/50	52/30	164/91
	Age	41 [17,76]	33 [16,60]	47 [18,84]	39 [16,84]
	DPI	14 [0,23]	2 [0,5]	17 [1,42]	8 [0,42]
	Good (%)	26 (67%)	95 (71%)	34 (41%)	155 (61%)
	Poor (%)	13 (33%)	39 (29%)	48 (59%)	100 (39%)
	GCS 13/14/15	2%/13%/85%	3%/15%/82%	7%/26%/67%	4%/18%/78%
	GOSE	6 [5, 7]	7 [5, 7]	6 [3, 7]	7 [3, 7]

Sex displayed as male (M)/female (F). Age at scan (*Age*), days post injury (*DPI*) sown as mean[min, max]. Patients dichotomised in groups with *good* (GOSE=8) or *poor* (GOSE<8) outcome, are displayed as absolute number and as percentage, indicating the portion of all patients. *GCS* indicates the proportion of patients with GCS scores equal 13, 14 or 15. *GOSE* shown only for poor outcome patients as median [min, max]

Models mentioned above were fitted to individual ROIs, which did not account for dependencies between ROIs. Hence, a GLM was fitted to predict outcome using the information of all ROIs from the pooled databases simultaneously.

Model #4:

$$f_\eta(Y) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Site_i + \beta_4 GCS_i + \beta_5 DPI_i + \alpha_{1-n} ROI_{i,1-n} + \varepsilon_i$$

with $ROI_{i,1-n}$ representing the regional information and n is the number of ROIs (e.g. 24 for TractSeg). Complexity of the model was neglected for this analysis. The variables $\beta_{0,\dots,5}$ and α_{1-n} are the parameters to be fitted.

3.3.7 Longitudinal Analysis

For the longitudinal comparison patients were only included if they had at least two MR scans (Table 3.3). The aim was to examine whether volumes or diffusion metrics change over time differently for patients with good or poor functional outcome. Therefore, regional metrics were expressed as a linear function of time after injury (i.e. DPI) in interaction with the dichotomised outcome. Since each patient was likely to have a different baseline to start

Table 3.3: Overview of Available Data for Longitudinal Analysis

	Category	Cambridge	Trondheim	Turku	All
Controls	# Subjects/Scans	42/47	83/206	30/51	155/304
	Sex (M/F)	23/19	50/33	14/16	87/68
	Age	39[18,78]	33[16,59]	49[22,91]	37[16,91]
Patients	# Subjects/Scans	41/172	136/377	87/174	264/723
	Sex (M/F)	29/12	87/49	56/31	172/92
	Age	42[17,74]	34[16,60]	47[18,84]	39[16,84]
	DPI (acute)	11[0,23]	2[0,5]	17[1,42]	8[0,42]
	DPI (3 months)	94[81,123]	96[80,122]	56[43,100]	93[43,123]
	DPI (6 months)	198[174,253]	-	223[151,299]	215[151,299]
	DPI (12 months)	381[357,431]	371[348,426]	394[336,491]	374[336,491]
	DPI (chronic)	736[681,798]	-	-	736[681,798]
	Good (%)	25(61%)	96(71%)	34(39%)	155(59%)
	Poor (%)	16(39%)	40(29%)	53(61%)	109(41%)
	GCS 13/14/15	2%/13%/85%	4%/15%/81%	9%/24%/69%	5%/17%/78%
	GOSE	6 [5, 7]	7 [5, 7]	6 [3, 7]	7 [3, 7]

Sex displayed as male (M)/female (F). Age at scan (*Age*), days post injury (*DPI*) sown in mean[min, max]. Patients dichotomised in groups with *good* (GOSE=8) or *poor* (GOSE<8) outcome, are displayed as absolute number and as percentage, indicating the portion of all patients. *GCS* indicates the proportion of patients with GCS scores equal 13, 14 or 15. *GOSE* shown only for poor outcome patients as median [min, max]

with, a linear mixed effect model was fitted to the data to allow a random intercept for each patient. Age, sex, acquisitions site as well as GCS were considered as confounding factors.

Model #5:

$$\mathrm{f}_\gamma(Y) = \beta_0 + \beta_1 Outcome_i * DPI_i + \beta_2 Age_i + \beta_3 Sex_i + \beta_4 Site_i + \beta_5 GCS_i + \beta_6 Outcome_i + \beta_7 DPI_i + \varepsilon_i$$

where $\beta_{0,\ldots,7}$ are the parameters to be fitted by the GLM, f_γ representing the log link function to fit a Gamma distribution and ε_i is the error unexplained by the model. The dependent variable (Y) was replaced with continuous volume or diffusion measurement of an individual ROIs, for which a separate GLM was fitted.

3.4 Results

3.4.1 Site-Specific Biases

Volumes. Initial results of the regression analysis of Z-scored volumes have disclosed that half of the examined ROIs (ten), demonstrated on average different volumes across centres. Among these were the whole cortical volume ($p_{fdr} = 0.0367$), the right precuneus ($p_{fdr} = 0.0496$), and in both hemispheres the nucleus accumbens (left: $p_{fdr} = 0.0223$, right: $p_{fdr} = 0.0035$), putamen (left: $p_{fdr} = 0.0041$, right: $p_{fdr} = 0.0041$), thalamus (left: $p_{fdr} = 0.0087$ right: $p_{fdr} = 0.0121$) and the anterior cingulum (left: $p_{fdr} = 0.0228$ right: $p_{fdr} = 0.0223$). Although the total brain volume was increased for Turku subjects ($p = 0.0496$), this was statistically not significant after FDR correction ($p_{fdr} = 0.0763$).
The number of ROIs with deviating volumes was reduced to three by normalising the regional volumes by the total brain volume for each subject individually before Z-scoring. Nonetheless, among the 19 selected ROIs (excluding total brain volume), control subjects from Turku showed significantly higher volumes for the right nucleus accumbens ($p_{fdr} = 0.0164$, $\beta_1 = 0.2770$) as well as the left and right putamen (left: $p_{fdr} = 0.0286$, $\beta_1 = 0.1192$, right: $p_{fdr} = 0.0329$, $\beta_1 = 0.2717$). Figure 3.2 shows a shift towards higher volumes in Turku controls compared to both other databases. Other regions with increased volumes in Turku controls were the left nucleus accumbens ($p = 0.0334$) and both the left and right anterior CG ($p = 0.0365$ and $p = 0.0373$, respectively), however, these trends were not statistically significant.

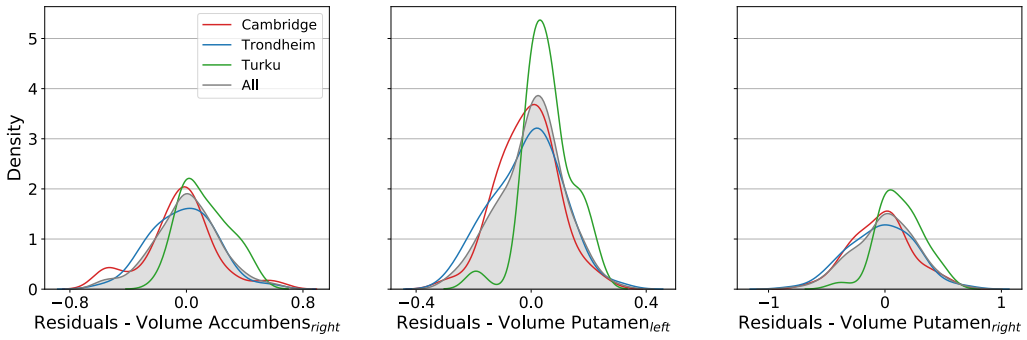


Figure 3.2: Examples of Regional Volume Differences Across Sites. The residuals of the GLM after correcting for age and sex differences, still showed a deviation between controls from Turku and the other two imaging sites within the accumbens and putamen. Area shaded in grey highlights the area under the distribution of all subjects.

Examining combinations of pairs of databases, showed no differences between Cambridge and Turku, but significant differences between Trondheim and the other two databases were present. Affected regions were WM and cortex overall as well as the caudate and thalamus

in both hemispheres ($p_{fdr} < 0.05$). Differences between imaging sites were most pronounced for Trondheim and Turku.

Fractional Anisotropy. When jointly analysing all three databases together, no significant differences were found for any of the 24 pre-selected TractSeg ROIs between the three sites (for all ROIs: $p > 0.4344$). The rostrum showed a trend of higher FA values for Turku controls compared to Cambridge controls ($p = 0.0414$, $\beta_1 = 0.1109$), however, this could not statistically be confirmed ($p_{fdr} = 0.7494$). Comparing Cambridge controls against Trondheim data showed no significant differences (for all ROIs: $p > 0.2909$). In contrast, Turku controls were found to have divergent FA values from Cambridge controls in the right ART ($p = 0.0285$, $\beta_1 = 0.0775$), rostrum ($p = 0.0458$, $\beta_1 = 0.1433$), right CST ($p = 0.0160$, $\beta_1 = 0.2263$) and right SLF_{III} ($p = 0.0499$, $\beta_1 = 0.0785$). However, none of the ROIs showed significant FA differences after FDR correction (for all ROIs: $p_{fdr} > 0.2200$). Similarly, the left ATR ($p = 0.0463$, $\beta_1 = 0.0683$), rostrum ($p = 0.0144$, $\beta_1 = 0.1215$) and genu ($p = 0.0395$, $\beta_1 = 0.0819$) demonstrated higher FA values for Turku than Trondheim, but differences were not statistically significant after correction for multiple comparisons (for all ROIs: $p_{fdr} > 0.3444$).
Figure 3.3 shows the distribution of the GLM residuals (accounted for age and sex) for the detected ROIs with FA variation across sites. While the distributions of Cambridge and Trondheim mostly align, FA metrics seemed slightly elevated for Turku controls.

Mean Diffusivity. None of the chosen TractSeg ROIs showed significantly different MD values across the sites in a joint analysis of all three databases ($p > 0.3742$). Similarly, combining Cambridge either with Trondheim or Turku data alone revealed no noticeable different regions (for all ROIs: $p > 0.2489$ and $p > 0.2552$, respectively). Only when leaving out Cambridge data, MD in the left ATR ($p = 0.03307$, $\beta_1 = -0.2133$) and right IFO ($p = 0.0426$, $\beta_1 = -0.2115$) was found to be slightly decreased in Turku with respect to Trondheim. Both findings were not statistically significant after FDR correction (ATR_{left}: $p_{fdr} = 0.4487$, IFO_{right}: $p_{fdr} = 0.4487$).

3.4.2 Acute Differences between Controls and Patients

Volumes. Volumes of ventricles overall ($p = 0.0298$, $\beta_1 = 0.1696$) and specifically the right lateral ventricle ($p = 0.0470$, $\beta_1 = 0.1622$) were found to be increased in patients with poor outcome compared to controls, however, this was not statistically significant after FDR correction. Likewise, the right anterior CG showed a decreased volume for both patients with good ($p = 0.0266$, $\beta_1 = -0.1087$) and poor outcome ($p = 0.0336$, $\beta_1 = -0.1192$) compared

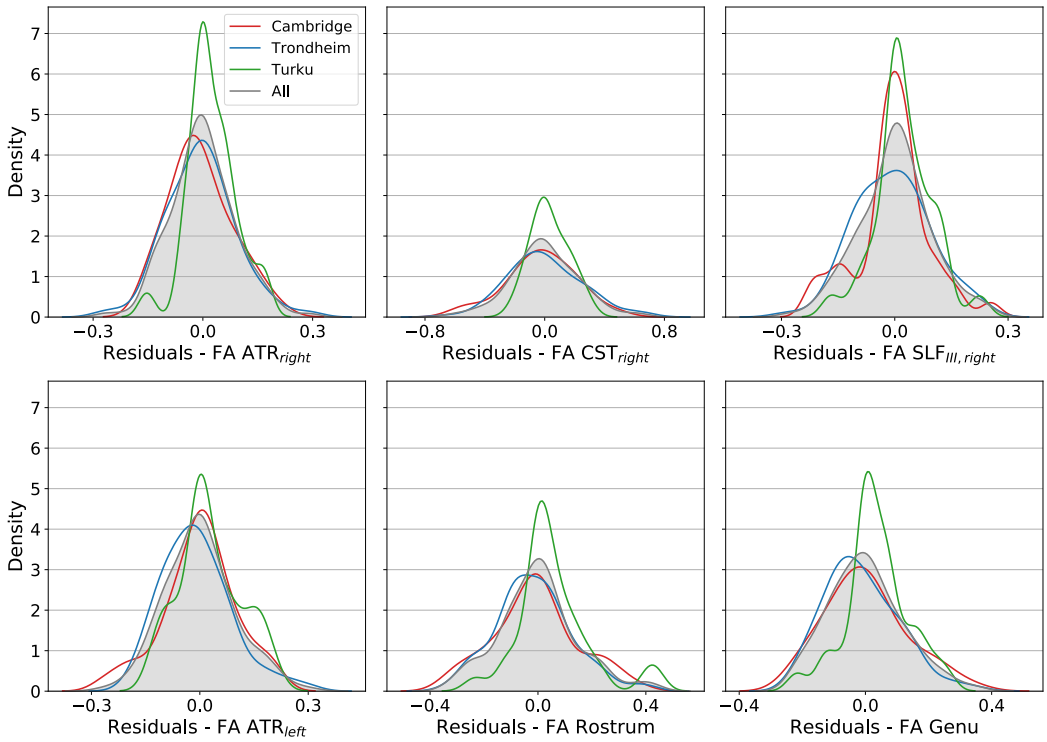


Figure 3.3: Regional FA Differences in Control Subjects. The residuals of the GLM after correcting for age and sex differences, still showed a deviation between controls from Turku and the other two imaging databases within the ATR, CST, SLF$_{III}$, rostrum and genu. Differences were not significant after FDR correction.

to controls, but neither of the differences was statistically significant after FDR correction. Analysis of individual data did not reveal any deviations between the three subject groups.

Fractional Anisotropy. A comparison of data from all three sites revealed 15 ROIs with significantly lower mean FA values in patients with poor outcome relative to controls. However, patients with good outcome did not show any significant deviation from subjects without head injury (Table 3.4). Considering only subjects scanned in Cambridge, FA values in both the left and right IFO ($p_{fdr} = 0.04717$ and $p_{fdr} = 0.0403$, respectively), as well as in the left UF ($p_{fdr} = 0.0471$) were significantly lower in the patient group with poor outcome. However, patients with good outcome did not statistically differ from controls. No significant FA differences were found between the three subject groups within the Trondheim database. In contrast, Turku patients with poor outcome demonstrated reduced FA values in 17 ROIs compared to controls (Table 3.5). Turku patients with good outcome were statistically indistinguishable to controls.

Analysing regional FA differences in the three databases individually and together revealed various patterns. For example, FA was reduced in IFO$_{left}$ for Cambridge patients with poor

outcome. This difference was not observed in the two other databases, but remained present in the joint analysis. On the other hand, regions such as the ILF_{left}, showed no significantly different FA values between subjects at individual sites, but revealed lowered FA values in patients with poor outcome in the combined databases. Cambridge patients with poor outcome had lower FA values in ILF_{left}, which, however, was deemed non-significant after FDR correction. The joint analysis allowed to recover this difference. In contrast, combining databases also diminished previously found differences. While FA value in $\text{SLF}_{I,left}$ was lower for Turku patients with poor outcome, this difference was not statistically present in the pooled dataset. Besides those changes, some regions showed decreased FA values across databases. For example the UF_{left} demonstrated significantly decreased FA values for patients with poor outcome in the Cambridge and Turku as well as the fused databases (Figure 3.4). From this analysis we can only infer changes when looking at different databases, but cannot say with certainty which differences between changes are true discrepancies between subjects.

Table 3.4: Overview of ROIs with Different FA in Acute Phase Across Sites

Cambridge & Trondheim & Turku					
	ROI	good - β_1	good - p_{fdr}	poor - β_1	poor - p_{fdr}
FA	ATR_{right}	-0.0130	0.9748	-0.0504	0.0271
	Rostrum	-0.0549	0.7731	-0.1142	0.0110
	Genu	-0.0292	0.9748	-0.0731	0.0271
	PMB	-0.0013	0.9748	-0.0984	0.0498
	CG_{left}	-0.0157	0.9748	-0.1366	0.0152
	CG_{right}	-0.0081	0.9748	-0.1260	0.0115
	IFO_{left}	-0.0243	0.9748	-0.0921	0.0271
	IFO_{right}	-0.0052	0.9748	-0.1079	0.0115
	ILF_{left}	-0.0035	0.9748	-0.0972	0.0271
	ILF_{right}	-0.0025	0.9748	-0.1243	0.0123
	$\text{SLF}_{I,right}$	-0.0300	0.9748	-0.0950	0.0271
	$\text{SLF}_{II,right}$	-0.0091	0.9748	-0.0625	0.0311
	UF_{left}	-0.0635	0.7731	-0.1184	0.0123
	UF_{right}	0.0252	0.9748	-0.0942	0.0271
	CC	-0.0067	0.9748	-0.0846	0.0311

Model #2 coefficients (β_1) and p-values after FDR correction (p_{fdr}) for patients with good and poor outcome.

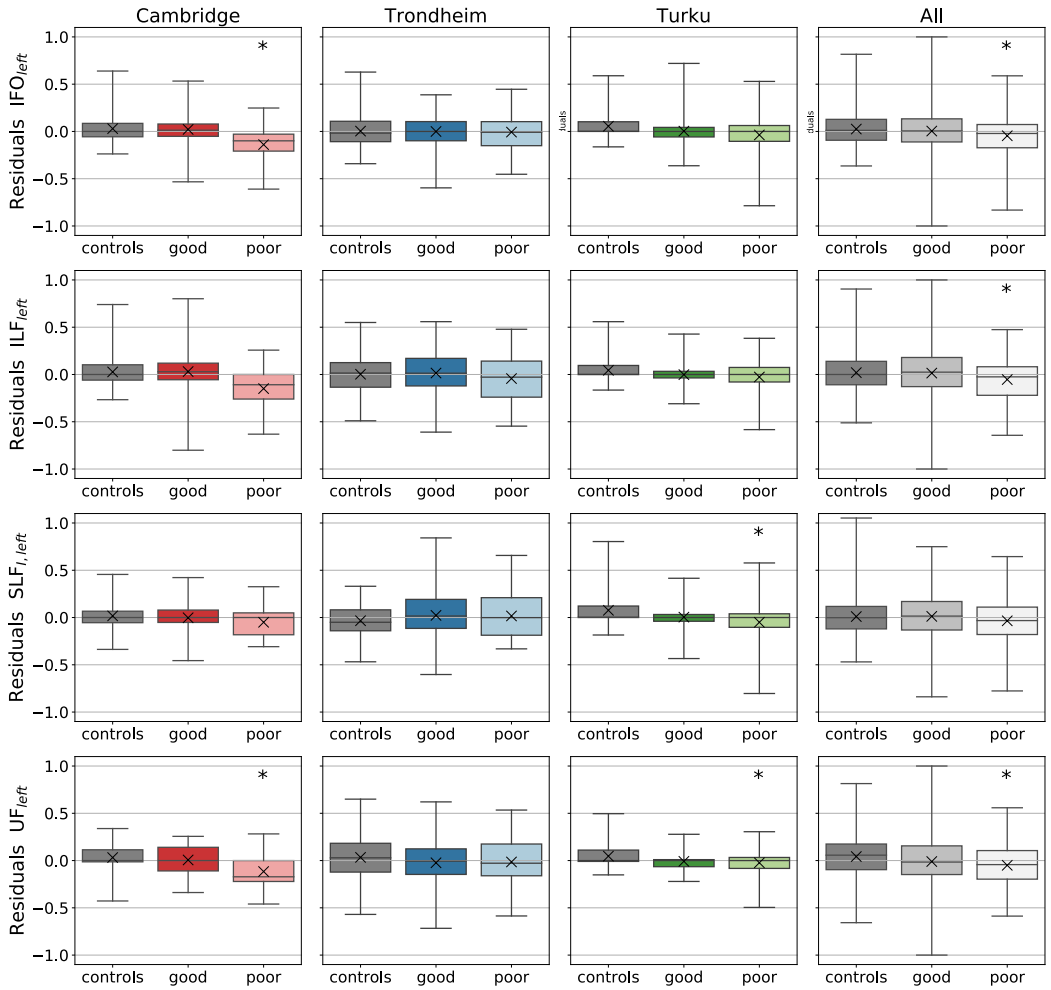


Figure 3.4: Examples of Regional FA Differences in Individual Databases and Joint Analysis After Correction for Confounding Factors. **IFO**$_{left}$**:** FA values were lowered in Cambridge data (*red*) and combined analysis (*All*). Trondheim (*blue*) and Turku (*green*) subject were indistinguishable within the particular database. **ILF**$_{left}$**:** FA values, despite lowered in poor Cambridge patients, were not significantly different when examining databases individually, however, joint analysis recovered this difference. **SLF**$_{I,left}$**:** FA was significantly reduced in patients with poor outcome in Turku, but not in Cambridge or Trondheim. In a joint analysis this difference was not present. **UF**$_{left}$: Mean FA values were decreased for patients with poor outcome in Cambridge and Turku as well as the combined analysis. Mean values are marked with a cross (×). Regions with significantly lowered FA are highlighted with an asterisk (*)

Table 3.5: Overview of ROIs with Different Diffusion Metric in Acute Phase

Cambridge					
	ROI	good - β_1	good - p_{fdr}	poor - β_1	poor - p_{fdr}
FA	IFO_{left}	0.0175	0.9752	-0.3066	0.0471
	IFO_{right}	0.0287	0.9752	-0.3370	0.0403
	UF_{left}	-0.0380	0.9752	-0.2717	0.0471
MD	Rostrum	0.0843	0.9143	0.5686	0.0203
	CST_{left}	-0.0024	0.9879	0.5374	0.0186
	IFO_{left}	-0.0767	0.9143	0.5188	0.0430
	ILF_{left}	0.0271	0.9572	0.4974	0.0296
	UF_{left}	0.1092	0.8390	0.4080	0.0203
Turku					
	ROI	good - β_1	good - p_{fdr}	poor - β_1	poor - p_{fdr}
FA	ATR_{right}	-0.0423	0.3783	-0.0724	0.0418
	Rostrum	-0.0566	0.4517	-0.1391	0.0331
	Genu	-0.0679	0.3783	-0.1298	0.0283
	AMB	-0.1126	0.3783	-0.1337	0.0319
	PMB	-0.2540	0.3590	-0.3795	0.0145
	CG_{left}	-0.2720	0.3590	-0.3341	0.0148
	CG_{right}	-0.1957	0.3590	-0.2713	0.0145
	IFO_{right}	-0.1005	0.3783	-0.1712	0.0283
	ILF_{right}	-0.2375	0.3590	-0.3246	0.0145
	$SLF_{I,left}$	-0.1466	0.3783	-0.2429	0.0222
	$SLF_{I,right}$	-0.1152	0.3783	-0.2037	0.0222
	$SLF_{II,left}$	-0.0949	0.3783	-0.1696	0.0331
	$SLF_{II,right}$	-0.0772	0.3783	-0.1508	0.0222
	$SLF_{III,left}$	-0.0747	0.3783	-0.1125	0.0319
	$SLF_{III,right}$	-0.0498	0.3783	-0.0880	0.0418
	UF_{left}	-0.1134	0.3783	-0.1327	0.0474
	CC	-0.1138	0.3783	-0.1990	0.0222
MD	ATR_{right}	0.1922	0.5885	0.2326	0.0385
	PMB	0.1367	0.5885	0.3304	0.0037
	ILF_{right}	0.0993	0.5885	0.1456	0.0496

Model #2 coefficients (β_1) and p-values after FDR correction (p_{fdr}) for patients with good and poor outcome.

Mean Diffusivity. None of the examined TractSeg ROIs showed different mean MD values between controls and patients with good outcome in the pooled dataset. Patients with poor outcome had a trend towards elevated MD values relative to controls within the rostrum ($p = 0.0222$, $\beta_1 = 0.1381$), the right ILF ($p = 0.0391$, $\beta_1 = 0.0780$) and the left UF ($p = 0.0234$, $\beta_1 = 0.0964$). However, those results were statistically insignificant (for all three ROIs: $p_{fdr} > 0.2805$). Analysing differences between subject groups within the Cambridge databases, uncovered increased MD values for patients with poor outcome in the rostrum ($p_{fdr} > 0.0203$, $\beta_1 = 0.5686$) as well as in the CST ($p_{fdr} = 0.0186$, $\beta_1 = 0.5374$), IFO ($p_{fdr} = 0.0430$, $\beta_1 = 0.5188$), ILF ($p_{fdr} = 0.0296$, $\beta_1 = 0.4974$) and UF ($p_{fdr} = 0.0203$, $\beta_1 = 0.4080$) in the left brain hemisphere. No differences were found for patients with good outcome (Table 3.5). Similarly to previous findings, MD values were statistically inseparable for the controls and patients within the Trondheim databases. The Turku database also demonstrated increased MD values in the ATR_{right} ($p_{fdr} = 0.0385$, $\beta_1 = 0.2326$), the PMB ($p_{fdr} = 0.0037$, $\beta_1 = 0.3304$) and the ILF_{right} ($p_{fdr} = 0.0496$, $\beta_1 = 0.1456$) for patients with poor outcome, but overall no MD differences for patients with good outcome (Table 3.5).

3.4.3 Prognostic Value of Acute MRI

Volumes. Although regression analysis indicated a volume difference for the right caudate between patients with good and poor outcome ($p = 0.0478$), none of the 19 selected MALP-EM ROIs were predictive of functional outcome several months post-injury. The left nucleus accumbens ($p = 0.0330$) and the right cerebellar WM ($p = 0.0132$) showed a potential predictive power within the Trondheim data, however, differences were not strong enough to remain statistically significant after FDR correction. Similarly, the right nucleus accumbens ($p = 0.0376$) and both caudates (left: $p = 0.0050$, right: $p = 0.0145$) in Turku data showed differences for both patient groups but differences did not remain after FDR correction. Regression analysis in Cambridge database alone did not allow to derive any results, as the variables used as covariates perfectly separated the small number of patients. Hence, the suggested regression model could not be fitted to infer differences between patients with good or poor outcome. Only when jointly analysing of Cambridge and Turku data while excluding Trondheim patients, volumes for the cortex ($p_{fdr} = 0.0065$) and the caudates (left: $p_{fdr} = 0.0049$, right: $p_{fdr} = 0.0037$) were predictive of functional outcome. Linear regression analysis of all selected MALP-EM ROIs (*Model#4*) did not reveal any regional volumes that were predictive of patient outcome.

Fractional Anisotropy. Considering all three databases together, FA in the right IFO fascicle ($p = 0.0310$), left ILF ($p = 0.0463$) and right UF ($p = 0.0177$) had the potential to predict patients with good and poor outcome. However, after FDR correction none of the regions showed a statistically significant impact on outcome prediction. Examining only Trondheim data did also not reveal any statistical significant FA differences between patient with different outcome. In Turku patients, the FA of the right CST ($p = 0.0203$) may be predictive for poor outcome, but this was not statistically significant after FDR correction. As before, Cambridge database included too few patients for the regression analysis. Noteworthy, the acquisition site was a strong predictor of outcome. In fact, when leaving out Trondheim patients, 14 ROIs showed a discriminative power ($p_{fdr} \leq 0.05$) to distinguish both patient groups. Among the three different centres, Trondheim scanned patients with the mildest TBI. This means it is hardest to differentiated patient groups. Including Trondheim patients may diminish the discrepancy between both patient groups that may be more prevalent in Cambridge and Turku data. This discrepancy was uncovered when excluding Trondheim patients, which previously have masked the subtle differences (hence the 14 ROIs). This suggested site-/cohort-specific biases. Other combinations of two databases did not show the same effect. Although it is difficult to say what effects are truly a characteristics of patients with good or poor outcome, this shows the heterogeneity of TBI patients. A GLM (*Model#4*) fitted to mean FA values from all ROIs simultaneously for the fused databases revealed the AMB ($p = 0.031$) and right UF ($p = 0.048$) to be predictive of patient outcome.

Mean Diffusivity. Similar to FA, none of the 24 TractSeg ROIs showed a predictive influence on patient outcome. This was observed in both the analysis of fused databases as well as in Trondheim and Turku individually. As before, sample size for Cambridge was too small to infer results. On the other hand, excluding Trondheim patients resulted again in an increase of relevant ROIs. Average values for MD were predictive of outcome in 17 TractSeg ROIs ($p_{fdr} \leq 0.05$). When fitting a linear regression model to all regional MD values for the pooled databases (*Model#4*) both ATRs (left: $p = 0.002$, right: $p = 0.026$) as well as the left cingulum ($p = 0.038$) and right IFO ($p = 0.010$) were predictive of patient outcome. Considering all ROIs simultaneously to predict outcome may be more sensitive to find outlier regions than analysing individual fibre tracts. With more information provided, the linear model seems to be fitted to the data differently to map the random variables (i.e. here MD values) to patient outcome. From this analysis it is not clear whether this is due to better or worse data fitting through the GLM.

3.4.4 Longitudinal Analysis

Volumes. After accounting for age, sex, site and injury severity only volumes of the right and left caudate demonstrated a minimal negative dependency on DPI (left: $p = 0.0132$, $\beta_1 < -0.0002$, right: $p = 0.0132$, $\beta_1 = -0.0002$). In other words, both regions showed a trend of progressing volume loss. The left cerebellar WM showed a positive association with time after injury ($p = 0.0044$, $\beta_1 = 0.0005$), indicating a higher volumes at a later stage. These findings were, however, not statistically significant after correction for multiple comparison ($p_{fdr} > 0.0831$). Fitting a regression model to all three databases together did not reveal any differences in regional volume loss over time between patients with good and poor outcome. Similarly, analysing Cambridge and Trondheim databases individually resulted in the same findings. A difference in temporal tissue atrophy between both outcome groups was found for the overall cortex in Turku patients ($p = 0.0428$, $\beta_1 = 0.001$), but this difference was not significant after FDR correction ($p_{fdr} = 0.6667$).

Fractional Anisotropy. A dependency of FA on time after injury was found for six TractSeg ROIs. The CST in both hemispheres (left: $p_{fdr} = 0.0297$, $\beta_1 = -0.0004$, right: $p_{fdr} = 0.0045$, $\beta_1 = -0.0006$) and the IFO ($p_{fdr} = 0.0014$, $\beta_1 = -0.0006$), ILF ($p_{fdr} = 0.0017$, $\beta_1 = -0.0006$) and SLF_I ($p_{fdr} = 0.0297$, $\beta_1 = -0.0004$) on the right side showed a significant negative correlation with DPI. This suggested progressively lower FA values in those regions after mTBI. On the other hand, the rostrum showed a significantly positive association with DPI ($p_{fdr} = 0.0017$, $\beta_1 = 0.0005$), indicating an increase in FA values over time. The right IFO ($p_{fdr} = 0.0490$, $\beta_6 = -0.4950$) and right UF ($p_{fdr} = 0.0490$, $\beta_6 = -0.4190$) showed a significantly stronger decreased of FA in patients with poor than with good outcome. Analysing the three databases individually revealed slightly negative trends for patients with poor outcome in the rostrum for the Trondheim cohort ($p = 0.0096$, $\beta_1 = -0.001$) and in the PMB for the Turku cohort ($p = 0.0323$, $\beta_1 = -0.002$). Differences could not be statistically verified after FDR correction ($p_{fdr} > 0.2295$).

Mean Diffusivity. Time post injury was found to also have an effect on regional MD. In particular, diffusion in the splenium ($p_{fdr} = 0.0049$, $\beta_1 = 0.0005$), CST (left: $p_{fdr} = 0.0069$, $\beta_1 = 0.0006$, right: $p_{fdr} = 0.0049$, $\beta_1 = 0.0006$), right IFO ($p_{fdr} = 0.0049$, $\beta_1 = 0.0005$), ILF (left: $p_{fdr} = 0.0313$, $\beta_1 = 0.0004$, right: $p_{fdr} = 0.0158$, $\beta_1 = 0.0005$) and right SLF_I ($p_{fdr} = 0.0204$, $\beta_1 = 0.0005$) showed a significant dependency on DPI. This implied an increase of MD values over time in those seven regions. In contrast, the rostrum ($p_{fdr} = 0.0234$, $\beta_1 = -0.0004$) and genu ($p_{fdr} = 0.0371$, $\beta_1 = -0.0004$) were associated with negative slopes

for DPI, indicating decreased MD. Although MD seemed raised at later stages, there were no significant differences between patients with good or poor outcome. This was also the case, when fitting an individual regression model to either Cambridge or Turku data. Trondheim patients with poor outcome by themselves demonstrated a negative slope for the left and right SLF_I (left: $p = 0.0411$,$\beta_1 = -0.001$, right: $p = 0.0121$,$\beta_1 = -0.001$), right SLF_{II} ($p = 0.0427$,$\beta_1 = -0.001$) and right SLF_{III} ($p = 0.0487$,$\beta_1 = -0.001$). These observations in the Trondheim database were not significant after multi-comparison correction ($p_{fdr} > 0.2893$).

3.5 Discussion

3.5.1 General Challenges in Database Comparability

Since TBI disease patterns are time-dependent [19, 43, 171, 273] the imaging time point needs to be considered for any analysis. However, independently designed studies may offer a different scan collection scheme that hampers pooling all data for a cross-sectional analysis. Among the databases presented here, for example, Trondheim acquired scans at three, but not six months post-injury. In contrast, Turku collected almost all of the follow-up MR scans at the six months time-window. Thus, both time-points - three and six months post-injury - cannot be examined in a joint analysis of all three databases. Furthermore, this deviation was also observed for other clinical assessments such as the GOSE scores. Both Trondheim and Turku collected GOSE scores during a patient's follow-up visit. Hence, for the prognostic evaluation of acute MR scans GOSE scores from three and six months had to be combined. This followed the assumption that the functional outcome is fairly stable between three and six months post injury (80% of the 37 Cambridge patients showed a stable GOSE score between both time points, see also Section 3.3.5).

Even if clinical and imaging data are available for the same time points, retrospective multi-centre studies are likely to combine databases for which images were acquired with different scanning parameters. Some variations can be rectified fairly easily, such as for example the selection of equal number of b_0 volumes. Other parameters may be more difficult to harmonise, and could have a more direct impact on image quality. For all three datasets only single-shell data ($b = 1000\,s/mm^2$) was chosen to apply the same tensor fitting model. Cambridge and Turku acquired approximately 60 directions on one shell, whereas, Trondheim only acquired DWI along 30 non-collinear gradient directions. While MD metrics have been reported to be more affected by variations in the b-values, FA was found to be sensitive to the number of diffusion gradient directions [15]. Hence, acquisition-/site-specific biases

were expected to be more reflected in FA, due to different angular resolutions, than in MD values. The fact that ROIs were observed to be more often significantly different between subject groups could be an indication of site- and population-specific biases still present despite harmonising metrics (Z-scoring) and considering site as a co-variable in the linear regression models. Future experiments could investigate the impact of choosing the same number of diffusion directions. For this data this would mean to artificially reduce the angular resolution for Cambridge and Turku data by selected a subset of 30 directions that are most similar to the 30 directions acquired for Trondheim data. However, the retrospectively selected 30 directions would not have been optimised to find the best spatial distribution of directions and might lead to suboptimal results. The effect of number of directions has been explored for Parkinson [206] and similar experiments could provide valuable insight for the design of future multi-centre studies in the context of TBI. Since there is no groundtruth available for this clinical dataset, simulation of such datasets could help to understand the impact of different number of directions (the reader is also referred to [279]).

3.5.2 Biases Across Imaging Sites

Despite Z-scoring volumes and considering acquisition site as a confounding factor in the regression analysis, a site-specific bias was detected for half the MALP-EM ROIs. Differences between Trondheim and Turku were most pronounced, which may reflect the age differences between the cohorts despite accounting for age in the linear regression model (average age was 33 and 51 for Trondheim and Turku, respectively).[7] While this bias could be cohort-specific (larger heads or younger age) this should be reduced by normalising to the average volume in control subjects at each site. Alternatively, Z-scoring itself may introduce a bias. Control subjects scanned in Turku showed the widest range of ages (i.e. 69 years, compared to Cambridge: 60 years, and Trondheim: 43 years). This is also reflected in the highest average absolute age difference between controls and the cohort mean age (i.e. 17 years, compared to Cambridge: 10 years, and Trondheim: 11 years). This higher variation in the Turku control subjects may lead to less accurate volume normalisation with respect to age, particularly, as tissue atrophy was reported to accelerate with increasing age [67]. This hypothesis seems supported by the drastic decrease in number of ROIs with site-specific differences for controls after normalising regional volumes to total brain volume (including CSF). Since both the left and right putamen persisted to have a different volume across sites, this may indicate a potential systematic difference in segmentation.

No difference in diffusion metrics were found between Cambridge and Trondheim controls, which suggests that differences were either not prevalent despite the difference in angular

[7] Future experiments could aim to statistically model non-linear age effects.

resolution, or that Z-scoring could account for some of the acquisition-induced variation. In contrast, Turku control subjects showed slighter elevated FA values in a few ROIs compared to Cambridge controls. This again may be due to a bias introduces via Z-scoring. Fewer deviations were observed for MD than for FA. This may be attributed to the use of single-shell data benefiting MD, which FA is more sensitive to the different angular resolutions across centres. Since none of the findings were statistically significant, subsequent experiments are likely to show indeed differences resulting from TBI pathology rather than acquisition-specific biases exclusively.

3.5.3 Differences between Patients in Comparison to Controls

Patients with poor outcome were found to have increased ventricle volumes compared to control subjects. Since ventricles did not show a site-specific bias in the previous analysis, this indicates potential TBI-related brain tissue atrophy. Patients with good outcome did not show a similar trend, thus, enlarged ventricles in the acute phase could hint a worse outcome a few month after the injury. Both patient groups demonstrated a volume decrease of the right anterior CG. All findings were, however, not statistically significant after FDR correction, which shows the subtlety of differences between controls and patients with different outcome.

More prominent were the differences seen for FA. Among the examined WM tracts, 15 ROIs showed decreased FA for patients with poor outcome during the acute phase. Since this number of regions exceeded the few regions for which potential site-specific differences between controls were found, reduced regional FA could indeed be an indication for poor outcome, especially, as lower FA values were not observed for patients with good outcome. The finding of FA values in the good outcome group similar to control subjects, may indicate a lack of injury, which could explain the better functional outcomes for that patient group three to six months post-injury. Alternatively, this could indicate a better brain plasticity that allows those patients to recover from TBI more easily. When analysing individual databases, Trondheim and Cambridge patients showed none or only a low number of ROIs with deviating FA. In contrast, Turku patients with poor outcome had decreased FA values compared to controls in 17 ROIs. These more prevalent FA changes may be connected to the initial severity of injury [172]. Turku patients showed the largest proportion of patients with lower GCS scores (i.e. GCS=13: 7%, GCS=14: 26% of all patients) and with poor outcome (59%). In addition, some Turku patients had a worse outcome (minimum GOSE=3) than the other two patient cohorts. The discrepancy between individual databases and the fact that the analyses that were either based on Turku data alone, or in combined analysis with Turku data, had a similar number of abnormal ROIs, suggests that the differences found in

the joint analysis were strongly driven by Turku patients. Some differences observed in individual databases were both weakened or strengthened by the fusion of all three databases. The joint analysis of all three databases showed that poor outcome patients followed a trend towards elevated MD values in some of the regions during the acute phase. Although these findings were not statistically significant, the higher MD values may mark a differences between patients that will recover and ones that show persistent symptoms later on. Examination of individual databases also underlined this. Cambridge and Turku patients with poor outcome showed significantly higher MD values. In contrast, Trondheim patients, who had generally less severe injuries, did not demonstrate any deviating MD values, regardless of their outcome. Overall, abnormalities in MD were less pronounced - showing differences in a lower number of ROIs - than those for FA. Disregarding that MD benefited from better rectification of sites-specific acquisition (i.e. number of b_0 volumes, also see Section 3.5.1), one reason for this could be the inclusion of acute scans over a wide time span post-injury. Mean diffusivity has been shown to pseudo-normalise during the semi-acute phase [19], which may balance out any abnormalities found in the hyper-acute phase. Subdividing the acute phase in smaller time increments could help to identify more transient changes in FA and MD [60].

3.5.4 Prognosis Based on Acute MRI

Although the volume of the right caudate showed some predictive influence for patient outcome, none of the examined structural volumes showed significant values for mTBI prognosis. Regions such as the caudate, nucleus accumbens or cerebellar WM had slightly different volumes between both patient groups in individual databases. However, these trends were not statistically significant, and only achieved significance as predictors of outcome when Trondheim data were excluded. The early time-point of acute data acquisition for Trondheim patients and the lack in prognostic value of volume difference may indicate that progressive volume loss, although induced by the injury, may take some time to become severe enough to be measured on a group-wise level.

Similar observations were made for FA, where some regions showed a discrepancy between patient groups, but were not strong enough to provide any power for outcome prediction. Outcome in Trondheim patients could not be distinguished by FA metrics, supporting the idea of low injury severity and FA differences on acute MR were not present or too subtle to be observed. Excluding Trondheim patients from the analysis resulted in a higher number of ROIs with different FA between patients. Combining Cambridge and Turku patients may enhance differences between patients by increasing power, but not diluting the magnitude of difference between groups. Acquisition parameters (e.g. number of diffusion directions)

were similar for both databases, and site-specific biases were accounted for by Z-scoring and using *site* as covariate. Moreover, the previous comparison of control subjects from the three different centres has shown only potential site-specific biases in FA for four ROIs (none were significant). Therefore, abnormal FA may indeed show patient differences related to TBI, rather than only site-specific biases. Nonetheless, it is likely that some centre-biases persist, and allow a statistical model to identify whether data originated from Turku or Cambridge. Since Turku patients had worse outcome than Cambridge patients, identifying the site could falsely seem to help to predict outcome more accurately.
Similarly, MD did not provide any significant prognostic value to differentiate patients based on outcome, unless Trondheim patients were excluded.

3.5.5 Longitudinal Findings in Mild TBI Patients

Both the progressive volume loss [98] and unchanged volumes [278] between early and follow-up scans have been reported for regions such as the nucleus accumbens, caudate, putamen or thalamus. In the mTBI cohort presented here, an ongoing tissue atrophy could be observed for GM regions such as the caudate in both hemispheres. Furthermore, WM volume in the left part of the cerebellum seemed increased with time after injury. However, time-dependent differences between patients with good or poor outcome were not found. One reason for the lack of differences could be that conventional MRI is not very sensitive to mTBI pathology [188]. Although loss of cerebral WM volume between three and 17 DPI was recently observed for mTBI patients compared to control subjects, the study did not examine volumetric changes between patients with different functional outcome [215]. Further investigation is needed to understand if WM volumes derived from T1w images could provide sufficient information to distinguish patient groups with different outcomes. Another reason could be the complex subject-dependent pathological evolution within different regions. Although the regression model estimated a different baseline per subject, a group-wise analysis might be to crude to detect patient-specific volume loss. To account for different patient trajectories of regional tissue atrophy, a model could be trained to fit a different slope and interception for each subject. However, the available data did not allow optimal convergence of such a model. Future experiments on a larger dataset might be more successful.
As expected, diffusion metrics were much more sensitive to capture longitudinal differences. Some of the examined regions showed a progressive decrease of FA over time indicating ongoing disease evolution. This is coherent with findings in previous studies (e.g. [19, 273]). In contrast, mean FA values in the rostrum increased with time after injury, which may be attributed to tissue recovery over time. The CC is one of the largest WM tracts allowing the communication between the two hemispheres of the brain. As anterior part of the CC,

the rostrum connects parts of the frontal lobes. These are often involved in TBI due to the prevalence of impact at the front of the head. An increase in FA suggests an elevated anisotropic diffusion, which possibly could indicate better more structured and well aligned fibre tracts. Hence, an increase in FA in the rostrum, may be associated with a stronger connection between the frontal lobes. One previous study reported initially decreased FA values in the internal capsule and the IFO at seven DPI, which recovered at one month after the injury. Increased FA values in both tracts were positively linked to better performances in cognitive information processing [273]. So, one can cautiously hypothesise that over time neural structures are enforced within the rostrum to strengthen the connection between frontal lobes to account for possible deficiencies in the frontal lobes due to TBI. However, more experiments would be needed to confirm such a hypothesis. Despite accounting for confounding factors, the longitudinal analysis presented here might not be completely free from site-specific factors. Comparing controls across sites, showed a trend for higher FA values in the rostrum for Turku than for Trondheim or Cambridge patients. This could have had an effect on the longitudinal analysis. However, patient data from later stages (12 months or chronic phase) were predominantly acquired at Trondheim and Cambridge. Consequently, their site-specific lower FA values from these centres would be more likely to support a decrease in rostrum FA values over time. Hence, the rostrum increase cannot directly be associated with site-specific effect, which may suggest a true recovery of FA values over time.

The longitudinal analysis also showed a positive association of DPI and MD in some regions, such as the CST, IFO or ILF. Increasingly higher MD values could indicate less structured WM with time after injury. The IFO tract passes from the frontal lobe radiating into the temporal and occipital lobes. The latter is the visual processing centre within the brain [212], and the temporal lobe is involved in visual memory as well as language comprehension [42]. Similarly, the ILF has been recognised to be involved in processing visual cues and its disruption has been linked to neuropsychological impairments of visual cognition [99]. So, deterioration of WM in the IFO and ILF could lead to problems in processing visual and verbal information. Indeed, differences in FA and MD between TBI patients and controls in IFO and ILF were observed to correlate with visual and verbal memory [23]. Again, rostrum was the exception with seemingly lower MD values at a later stage. Despite the sensitivity of diffusion metrics to pathophysiology in TBI, neither FA or MD demonstrated greater temporal changes for patients with poor than with good outcome. Diffusion may change differently for each patient following an individual trajectory [268, 273], which makes it more difficult to find group-wise differences.

3.5.6 Heterogeneity of the TBI Cohort

Apart from site-specific confounding factors, the heterogeneity of subjects suffering from mTBI makes diagnosis/prognosis difficult [75, 102] and poses a challenge for any model to fit the data. In theory, patients could have the same GOSE score and similar clinical characteristics, but different regions that are affected by the TBI. A model based on group-wise analysis may not be flexible enough to fit to this complex data distribution. Including many more patients possibly enhances the changes of two subjects show similar patterns, which however, is uncertain due to the high variation of observed pathologies after TBI. Nonetheless, expanding the database would enable application of more data driven machine learning algorithms such as neural networks. However, these are not necessarily a guarantee for better performance [89]. Moreover, inference of valuable knowledge for clinical research from neural networks is more difficult, since it is not yet fully understood how these algorithms draw their conclusion [79]. Besides this, extending the database would incorporate data from many different sites, which comes with the drawback of an increased variation of non-diseases related biases. Consequently, the solution to these issue may lie in building models which, rather than targeting group differences, focus on anomaly detection for single-subjects. Such an approach based on TractSeg ROIs has been suggested for paediatric WM analysis [34]. Future experiments could investigate its application to TBI subjects.

3.6 Chapter Summary

Traumatic brain injuries are complex and show different pathological patterns closely linked to the time after injury. While most studies have been based on smaller single-site datasets, this chapter presented results from a multi-centre analysis including three mTBI databases. Experiments focused on imaging features, such as regional volumes and diffusion, derived from acute MRI as well as longitudinal changes. Hereby, differences between acquisition sites were accounted for via Z-scoring IDPs and including site as confounding factor for the regression analysis. Patients with good outcome seemed to display features similar to control subjects, whereas, patients with poor outcome showed enlarged ventricles and decreased FA. Nonetheless, patients with different outcome could not easily be differentiated based on their acute MR scans. Although longitudinal analysis revealed progressive tissue atrophy in the caudate, no differences were found for patients with good or poor outcome. Diffusion metrics displayed a stronger dependency on DPI, with more decreased FA or increased MD. However, patients with different outcome showed no differences in diffusion changes. The individual databases included mTBI cohorts of different severity and age, which also influenced the joint analysis. Acquisition-specific biases may have had an impact on enhancing differences

between subjects groups, but cannot be accounted for all observed variation.

Chapter 4

Reproducibility of MRI Metrics

4.1 Introduction

For many neuroimaging studies, features are derived from the MRI scans and compared for different groups of interest. This could involve volumetric changes over time or divergence between control and patient groups. Furthermore, mean intensity values of images within anatomical regions are often calculated to quantify differences between a patient cohort and healthy volunteers. Such findings can only be interpreted correctly if the applied methods, used to extract image derived features, are robust. With well validated tools this assumption usually holds true for MR images that were acquired under the same protocol on the exact same scanner. However, tools and measurements could be biased when considering multi-centre data. This could be especially problematic for the comparison of signal intensities across centres, since voxel values in common MRI sequences do not reflect quantitative absolute measurements. Different acquisition protocols or imaging equipment could lead to inherent biases that affect the performance of neuroimaging tools and with it the imaging derived results. To gain a better understanding of inter-scanner variability, this chapter first summarises related work and proceeds then to examine differences in brain parcellation and signal magnitude for datasets of various complexity.

4.1.1 Variability of Anatomical Brain Segmentation

While a number of studies have analysed the accuracy between automated brain region segmentations, only few have investigated the magnitude of inter-scanner variability. Jovicich et al. [120] examined the impact of acquisition variables on FreeSurfer's brain parcellation. The variability of major brain volumes (hippocampus, thalamus, caudate, putamen, lateral ventricles and total intracranial volume) was measured to be less than 4.3% on MR im-

ages collected on the same scanner within a few days. The reproducibility error was higher for smaller anatomical structures such as pallidum, amygdala and inferior lateral ventricle. Differences in bias field correction, MRI sequences, scanner updates and brain extraction affected the region volumes only minimally. However, a bias was identified when comparing inter-vendor data (Siemens vs. GE) and different field strengths (1.5T vs. 3T). Image quality factors (e.g. SNR) were found to have a strong influence on the segmentation outcome. Similar results were observed when subdividing T1w images with the *topology-preserving anatomy-driven segmentation* algorithm into nine cranial structures (cerebral GM and WM, cerebellar GM and WM, brainstem, caudate, putamen, thalamus, and ventricular CSF). Scan-rescan volume variation was less than 5% for all structures. The reproducibility of structural scans was found to be higher than for imaging modalities aiming to measure physiological quantities (e.g. DTI, functional MRI) [147]. A comparative study of hippocampal volume change has reported more coherent findings for automated segmentation than for manual annotations. Noteworthy is that FreeSurfer provided a higher reproducibility than manual segmentation or volume estimation with FSL FIRST. This was, however, only achieved when failed cases had been excluded after visual inspection [183]. Another recent study also explored the consistency of automated segmentation of the hypocampus in comparison to a manual approach. The intra-scanner variability for FreeSurfer volume estimation was less than 2% and comparable to that of manual annotations. Automatically derived hypocampal volumes were systematically different for Siemens and Philips scanners, resulting in a higher inter-vendor variation (*coefficient of variation* [CV], $CV = 4.4\%$) than that observed for manual region delineations ($CV = 2.6\%$) [49].

4.1.2 Reproducibility of Diffusion Magnetic Resonance Imaging

Diffusion tensor imaging metrics have been reported to show good scan-rescan repeatability on the same scanner regardless of vendor and scanner models [121], however, significant differences could be observed when comparing diffusion data across scanners. Various studies measured the intra- and inter-scanner variability for DTI metrics for different datasets and experiment setups. The following paragraphs aim to provide a non-exhaustive overview. Hereby, the focus lies on human brain imaging data analysing FA and MD metrics. The studies were roughly categorised according to whether their experimental design incorporated rescans on the same scanners, travelling volunteers or both.

Scan-Rescan Studies. To evaluate intra-scanner variation of DTI scans, Marenco et al. [168] scanned 20 subjects twice on one scanner and compared FA and MD measurements within 14 manually drawn regions. Variability was measured by the CV, which was defined

as the ratio of standard deviation and mean (see Equation 4.1). While nine out of 14 ROIs had a CV below 10%, there were noticeable regional differences, with CV scores ranging between 2.5% and 20.5% for FA and from 2.1% to 6.2% for MD. Highest CV for FA was found in the cerebellar cortex and frontal GM, and lowest was seen in the CC. The CV for MD was highest in the peduncles and lowest in the insula. Overall, MD showed more precise reproducibility (lower CV) than FA.

A later study, conducted by Lemkaddem et al. [154], compared intra- and inter-scanner variability of DTI scans on four different scanners. Two of them were located at the same site and shared the same configurations, hence thereafter referred to as *twin-scanners*. Although the intra-scanner measurements showed a lower variability than the inter-twin-scanner, overall no significant differences were observed. Intra- and inter-twin-scanner experiments revealed low CV for both region-based ($CV \approx 1.0\% - 4.2\%$) and tract-based ($CV < 3.0\%$) analysis. Comparing images across scanners showed a substantially higher variability with CV ranging between 4.0% and 8.9% for FA and from 4.2% to 10.0% for MD measurements. Some ROIs and tracts showed significant differences across the two scanners for MD but not for FA maps. The DTI metric deviation in ROIs and tracts was found to be larger between the two Siemens scanner models (Trio vs. Verio) than between scanners of the same model (both Trio), despite the unequal number of radiofrequency head coil channels for the latter. Noteworthy, data acquired with the lower number of channels exhibited a two fold reduction of SNR. Since, FA metrics have been shown to be sensitive to SNR [66], the characteristics of a head coil array may be an important factor to consider for multi-centre studies.

Vollmar and colleagues [256] analysed the reproducibility of DTI metrics in nine healthy subjects scanned twice on two different, but identical scanners. The variability of FA measurements was assessed both in the whole brain and in three manually outlined ROIs (splenium of CC, left frontal WM and left UF). The CV for FA within site ranged from 0.8% to 3.0%, and slightly increased across sites ranging from 1.0% to 4.1%. Highest variance was found in the smallest regions (left UF), and lowest in the whole brain average FA. Overall variation was lower in WM tracts (<5%) then in GM regions (10-15%). It has been mentioned that non-linear image coregistration improved reproducibility metrics in comparison to affine coregistration [256].

Another large multi-centre study [283] included 27 scanners (two vendors: Siemens & GE; six scanner models; seven software versions). Except for one scanner, MR data were collected with a harmonised single-shell DWI protocol (64 gradient directions, $b = 700\,s/mm^2$). To assess the variation across centres, a single and for each scanner different healthy volunteer was scanned twice.[1] The CV was computed for WM masks within single scanners and across

[1] one subject per centre, scanned twice: 27 subjects and 54 scans in total

scanners of the same vendor. This revealed no statistical deviation between intra-scanner images, however, significant differences were found for FA metrics across vendors. On average, FA and MD values were both lower on GE than on Siemens scanners. Cross-scanner variability was observed to be higher for FA ($CV = 5.7\%$) than for MD ($CV = 2.9\%$) [283]. This study showed overall good repeatability of DTI metrics across different scanners, but the results were only based on a single subject per scanner and were restricted to whole WM analysis. This may underestimate non-linear regional variability.

Travelling Head Studies. While studies with multiple scans and rescans on one scanner can confirm the reproducibility of DTI metrics at one site, they do not provide direct comparability between centres. Therefore, some studies have focused on acquiring data for the same subjects at different imaging sites and/or scanners. One of those [241] examined the comparability of DTI metrics in a multi-centre setup, involving 16 different scanners at 12 different imaging sites. Apart from one scanner (GE), all scanners were manufactured by Siemens, including various models, software versions and field strength (1.5T and 3T). The analysis of reproducibility was based on one healthy subject only[2]. For all scanners single-shell DWI data ($b = 0, 1000\, s/mm^2$) was acquired, however, with a different number of diffusion-weighting gradient directions (either 12 or 30). To measure the differences across scanners, the CV was computed in manually placed ROIs as well as for automated TBSS and deformation based analysis. For the latter, FA maps were affinely coregistered to T1w images, which were then spatially normalised, such that FA maps could be warped to template space. This revealed a variation of FA with a mean CV of 14% for TBSS and 29% for the deformation based analysis. The variation found was similar to the one observed between 12 healthy subjects and 26 Alzheimer's disease patients scanned at a single imaging site. Such a comparability shows the need to account for inter-scanner variances in clinical multi-centre studies. The magnitude of FA variation was only decreased marginally when choosing a subset of similar scanners and acquisition parameters. This could indicate that harmonised scanning protocols and identical scanners only partly diminish inter-scanner divergence. Interestingly, increased variability was observed for less organised fibre tracts, such as the fornix and SLF (in contrast to splenium). Despite the insight into inter-scanner DTI variability, findings are difficult to generalise as they are exclusively based on a single travelling subject. The impact on MD measurements was not analysed.
Slightly more elaborate was a study that included two travelling volunteers. These were scanned on five different scanners (three Siemens Trio and two GE Signa) with a harmonised DTI protocol including 33 directions. Mean values of FA and MD were evaluated within

[2] and a phantom, which will not be discussed here

16 manually delineated GM and WM ROIs. The intra-scanner CV within the WM was slightly lower (8.7%) than that across all five scanners (9.1%). Although a similar trend was found for MD, the relative discrepancy between intra- and inter-scanner CV (2.2% and 4.8%, respectively) was stronger. Overall, MD showed less variability than FA. Comparing regional DTI metrics across scanners revealed high concordance correlation coefficients for both FA (0.96) and MD (0.88). This high correlation of inter-scanner data indicates a good comparability of DTI metrics obtained with same imaging protocols, however, the low number of subjects may hamper generalisability [72].

For many studies, acquiring scans for one or several volunteers was not the main aim under investigation, but rather the methods for assessing the variability between scanners for a larger patient study. For example, in the case of the TRACK-TBI study, diffusion MR images were acquired for one travelling volunteer on 13 scanners at 11 different centres. Inter-scanner variability of DTI metrics was assessed by comparing intensities of the whole WM skeleton obtained from the standard FSL TBSS pipeline. This revealed small regional differences for the 14 selected fibre tracts, with CV ranging between 2.1% and 7.8% for FA and from 2.6% to 7.0% for MD. The globally measured CV values were below 5%, with MD exhibiting a lower variation than FA ($CV = 2.4\%$ and $CV = 4.2\%$, respectively). Moreover, DTI measurements were found to be more similar across imaging sites for large central fibre tracts (e.g. CC) than in small regions at the WM periphery (e.g. UF) [193].

Multi-Centre & Multi-Scan Studies. More advanced are studies that scanned several subjects repeatedly on all involved scanners, with the main aim to measure variability of DTI metrics in larger multi-centre data. For example, Magnotta and colleagues [164] scanned five healthy controls multiple times at eight imaging sites. Four DWI scans were acquired with a protocol that complied with the standard sequence for the employed scanner's vendor (30 directions on Siemens scanners or 32 directions on Philips scanners). Two further scans were collected with a semi-harmonised sequence with 71 directions but varying repetition times. All scans were single-shell data ($b = 1000\,s/mm^2$). Reproducibility was measured via CV within six selected ROIs from the Talairach[3] atlas [236] (cerebrum, frontal lobe, temporal lobe, parietal lobe, occipital lobe, and subcortical regions). Significant differences, assessed via mixed-effects model analysis, were found for inter-vendor comparison of DTI metrics. Mean FA values for Siemens scanners were slightly lower that for Philips scanners. Contrary, mean MD values were lower for Philips than for Siemens scans. While no differences were found for MD between the scanning protocols, the mean FA based on 71 directions was observed to be slightly lower than that calculated from the 30 or 32 directions.

[3] http://www.talairach.org

For most sites, the intra-scanner variability was below 1.0% for all DTI metrics. Coefficients of variation were slightly increased for inter-site comparison ($CV \sim 1.0\% - 3.0\%$), whereas FA showed marginally higher variability than MD. Variation was similar across regions, however, slightly elevated for all metrics within the occipital lobe. The mean CV for 30 or 32 direction images was lower (1.8%) than than for 71 direction scans (2.2%). Images were also denoised via median filters, which resulted in decreased CV scores. Variation scores were lower in Siemens than in Philips data for the 71 directions protocol, but the opposite was observed for 30 or 32 DWI scans. Mixed-effect model analysis suggested a significant impact on variability between sites through due to protocol type, vendor and median filtering [164]. In a recent study, Tong et al. [245] analysed the fibre-tract density on a voxel- and region-wise level and structural connectome for three travelling volunteers across eight imaging centres. The same multi-shell (b = 1000, 2000, 3000 s/mm^2) diffusion MRI scanning protocol was implemented on each of the similar scanners[4] (Siemens Prisma) to create ideal conditions of inter-scanner comparability. To form a baseline for intra-centre variability and investigate scan reproducibility on one scanner, the same subjects were scanned thrice at each site. Besides comparing the full DWI acquisition, all different combinations of shells and single-shell data were examined. A fair reproducibility of fibre-track density within and across centres ($CV < 15.0\%$) was observed. However, for all single- and multi-shell combinations the intra-centre variability was lower (average $CV = 10.4\%$) than for data across centres (average $CV = 11.3\%$). A closer look at tissue compartments revealed lower CVs for intra- than inter-scanner data at WM-GM boundaries ($CV_{intra} = 16.7\%$ vs. $CV_{inter} = 19.9\%$) and pure WM regions ($CV_{intra} = 6.6\%$ vs. $CV_{inter} = 8.0\%$). Interestingly, regions with more complex directional fibres (e.g. crossing fibres) showed generally higher variances than regions with single-directional fibres, which is consistent with the previously mentioned finding from Teipel et al. [241]. Reproducibility of links in the structural connectome was higher for intra-centre data, however, was also dependent on the number of considered regions. Finer parcellations resulted in less reproducible connections [245].

Besides considering the influence of scanners, the variability for different sequences has been systematically assessed as well. Eight healthy subjects were scanned with three diffusion imaging protocols (both with a different set of six gradient directions, and b-values of 1044 and 1034 s/mm^2) on two different scanners (Siemens & Philips). Additionally four healthy volunteers underwent diffusion imaging twice on the same scanner to quantify intra-scanner variability. This was measured by CV of histogram metrics within the whole brain excluding CSF. Among the derived quantities were the mean as well as the position and relative height of the peak of the histogram. Inter-sequence MD metrics exhibited less variation

[4] Here: inter-scanner and inter-centre are equivalent

($CV = 1.7\% - 5.6\%$) than FA measurements ($CV = 5.5\% - 7.3\%$). Variability for scans and rescans were comparable to inter-sequence CV scores [31].

Longitudinal Reliability. A different approach was pursued by Hawaco et al. [95] where four healthy subjects underwent annual DWI on five different scanners at three sites repeatedly over three years. Eventually, this longitudinal dataset included 27 MRI sessions acquired with a semi-harmonised scanning protocol (60 gradient directions, $b = 1000\, s/mm^2$, but varying TR). After spatial normalisation, mean FA values in 63 ROIs from the *Johns Hopkins University* (JHU) WM atlas [181] were computed. Estimating variability with mixed effect models highlighted a significant cross-scanner difference for FA. Furthermore, the aim was to investigate whether DTI metrics are reliable descriptors for individual subjects. Therefore, present inter-scanner effects were corrected for by regressing those from each ROI and a hierarchical clustering method was employed to classify subjects based on their image derived metrics. The high accuracy achieved, indicated a good reliability of FA metrics.

4.1.3 Overview & Aims

A number of studies [49, 120, 147, 183] have found that parcellating anatomical regions on structural MRI is fairly robust for data collected on different scanners. Automated region segmentation (i.e. FreeSurfer) has been reported to be equal or superior to manually outlined ROIs of inter-vendor data, but a bias between Siemens and Philips scanners has been observed [164].

A common finding of all DTI reproducibility studies was the lower variation for MD than for FA metrics. More specifically, a higher variability has been reported for FA in GM than in WM tissue regions [256]. Inter-scanner variability for different but identical scanners was low. However, statistical discrepancies for DTI metrics were observed for different scanner models from the same vendor as well for inter-vendor data. Although most studies report similar trends, a direct comparison is hampered, as different approaches were use to measure variability. Often analysis was performed on basis of manually delineated regions, which usually did not cover the full extent of WM fibre tracts. Despite offering a high precision of ROI segmentation, results become less reproducible when using non-automated brain parcellation. Furthermore, some studies were based on a low number subjects (1-2), which challenges the generalisability of those observations.

The objectives of this chapter is the evaluation of inter-scanner robustness of novel brain parcellation tools (i.e. MALP-EM [151] and TractSeg [263]) and measure the variability in DTI metrics for datasets with various complexity. Previous studies have restricted their

analysis on single subjects or small cohorts ($n \leq 3$), which limits the generalisability of any findings. Here the robustness of brain parcellations was measured in healthy subjects that were scanned twice on two scanners (T1w: $n = 12$, DWI: $n = 6$). This allowed the direct caparison of both intra- and inter-scanner variation of automated T1w and DWI parcellation as well as the variability of FA and MD metrics. Besides scanner differences, the influence of DTI parameter are examined in a cohort of healthy subjects ($n = 16$) scanned multiple times on the same scanner, with varying DWI acquisition protocol. Eventually, FA and MD variation was examined for a multi-centre study including nine physically different scanners. The chapter aims to provide a better insight in variation of MR data as a result of either different scanners, changing imaging protocols or both.

4.2 Data & Methods

4.2.1 Databases

Scan-Rescan Database As introduced previously (Section 2.2), this database included 12 subjects each scanned twice on each of two scanners. Six out of those subjects also underwent DWI. Acquisition protocols were mostly harmonised.

Multi-Acquisition Database For 16 subjects (nine female, seven male, age = 25.8±5.4), five DWI scans were acquired within the same scan session with identical acquisition parameters, but with different number of b-values. A total of 60 volumes with non-collinear gradient direction, evenly distributed across one to five shells, were collected. These schemes included a scan consisting of five shells (b = 250, 450, 700, 950 and 1200 s/mm^2) with each twelve directions (DTI 12×5), another included four b-values (b = 300, 600, 900 and 1200 s/mm^2) with each 15 directions (DTI 15×4). Alongside these, a scan with three b-factors (b = 400, 800 and 1200 s/mm^2) with each 20 directions (DTI 20×3) and a scan incorporating two b-values (b = 600 and 1200 s/mm^2) with each 30 directions (DTI 30×2) were collected. In addition to the multi-shell data a single-shell scan (b = 1200 s/mm^2) with 60 directions (DTI 60×1) was acquired. For each DWI scan 13 non-diffusion weighted volumes (b_0) were obtained, whereas one was always in the beginning and the remaining twelve were equally distributed between volumes of different shells and the very first b_0 volume (this means for single-shell data, all b_0 volumes are concatenated at the very beginning of the scan). Other parameters, such as $TR = 8000\,ms$, $TE = 93\,ms$, flip angle of 90° and pixel bandwidth of 1628 Hz/Px, were shared across the different DWI schemes. In addition one T1w images was collected ($TR = 2250\,ms$, $TE = 2.9\,ms$, flip angle= 9° and pixel band-

width of 230 Hz/Px). All images were acquired on one Siemens Trio scanner.

Multi-Centre Database The general MRI parameters for the CENTER-TBI data have been described previously (Section 2.2). To assess reproducibility healthy subjects were only included if they underwent a rescan with the same number of directions (hence, excluding three centres: *a72b20*, *ac3478* & *cb4e52*). Another centre (*8effee*) was excluded as only four scans from two subjects were left, which was deemed to be too low for this experiment. A further centre (*fe5dbb*) was excluded as scan and rescan had been processed with and without extra b_0 volume, respectively, as the additional b_0 scan with opposite phase encoding direction was not always available.

4.2.2 Experiment Setup

Reproducibility of brain parcellation and tissue segmentation via MALP-EM, TractSeg and JHU atlas projection was assessed. For this purpose, the *Scan-Rescan* dataset was used, as it allowed the direct comparison of multiple scans of healthy controls (12 for T1w, and 6 for DWI) on the same and a different scanner (Prisma & Trio). The harmonised MRI protocol and the paired scans on both scanners make this a highly controlled dataset. Mean intensities of FA and MD within TractSeg and JHU ROIs were compared to intra- and inter-scanner variation. The Multi-Acquisition database was used to examine the impact of different scanning parameters on FA and MD metrics as well as WM parcellation. Same principles were applied to the DWI data of the less well-matched multi-site CENTER-TBI database. For this, all healthy subjects, scanned twice, were selected and curated via the diffusion pipeline's QC metrics. These were then compared across centres to highlight any site-specific biases. Eventually, FA and MD variability was compared in different regions in native as well as template space.
Images were all pre-processed with the same structural and diffusion pipelines as outlined in the previous chapter (Chapter 2). The following sections describe the experimental setup in more detail.

Reproducibility of T1w Brain Parcellation Since MALP-EM is driven by image registration and ROI refinement through expectation maximisation, its parcellation is dependent on image intensities. While smaller errors of falsely deformed atlas regions might be cancelled out by fusing information from several atlases, it is not entirely clear how much impact differences in the image acquisition have on the region refinement. The examination of the robustness of MALP-EM was subdivided in three experiments.
Firstly, MALP-EM was applied to the 48 T1w scans of the Scan-Rescan database (12 sub-

jects, two scanners with each two sessions). For this the individual brain mask calculated via antsBrainExtraction was used. Besides the comparison of tissue and region parcellation, differences across scanners for relative region volumes were compared (relative volumes were computed by dividing ROI volumes by total brain volume).

Secondly, intensities of T1w images were first matched across scanners before re-applying MALP-EM. For this, T1w images collected on the Prisma scanner were matched to the mean intensity profile of the Trio T1w images with the recently suggested NDFlow[5] algorithm. This was chosen as it has been shown to be superior to standard intensity scaling. In brief, Dirichlet process Gaussian mixture models are fitted to the histograms of intensities found in the images to be matched. This allows to adapt the number of components to fit the histograms best, rather than being confined to a predefined, fixed number (as it would be the case for Gaussian mixture models). After an initial affine alignment of the histograms (similar to scaling image intensities), a non-linear transformation between histograms is estimated. This is done by considering the non-linear problem from the point of view of fluid mechanics: Particles, equating the source histogram's support points, follow individual trajectories to 'flow' to the target histogram. This forms a velocity field (the non-linear warp between histograms) that can be solved via ordinary differential equations [30]. For all scans the exact same brain masks as before were used, and since intensities for Trio scans were not adjusted, the previous MALP-EM results were reused. Another attempt for intensity harmonisation matched all scans from both Prisma and Trio scanners to the intensity space of a study-independent database (i.e. Cam-CAN, see Section 2.2).

Thirdly, for each subject separately all four T1w scans were rigidly coregistered to a common subject-specific template (NCC between registered images and subject-specific templates: $NCC = 0.988 \pm 0.004$). Corresponding ANTS brain masks were projected to this template space and fused by averaging them, thresholding the output at 0.5 and finally binarising the image to a common brain mask for all four scans. This was then used to re-apply MALP-EM on the coregistered T1w scans.

While the first experiment provides results from using a standard prepossessing approach, both other experiments aim to eliminate the impact factor of intensity shifts or skull stripping. Brain vessels and CSF (that is all CSF that is not in ventricles) were observed to be insufficiently segmented. Therefore, these regions (classified as "other" by MALP-EM) were excluded from any statistical analysis, reducing the number of ROIs for comparison from 138 to 133.

Reproducibility of DWI Parcellation & DTI Metrics Analogous to anatomical brain

[5] https://github.com/dccastro/NDFlow

scan parcellation, WM parcellation was examined on the *Scan-Rescan* database. This included six subjects with two DWI scans each on both scanners. For this reproducibility analysis two different approaches were chosen. Firstly, the brain was parcellated into 72 regions by applying TractSeg [263], a model based prediction algorithm. Secondly, 20 regions of the *Johns Hopkins University* (JHU) WM tractography atlas were projected to the DWI scan space via non-linear image registration. Volumes and weighted mean values for TractSeg were automatically computed by applying the diffusion pipeline to the DWI scans, hence no further processing was required. For JHU tract projection, each individual FA map was spatially normalised with antsRegistration to the JHU FA atlas as provided by FSL (JHU-ICBM-FA-1mm). The 20 JHU fibre tract probability maps were then backprojected from template to native space, thresholded at 25% and binarised to compute the volumes. A threshold of 25% was used as such binarised maps are also provided by FSL in MNI space, hence were considered as commonly used threshold.[6] Weighted mean values of FA and MD were computed before binarising projected tract probability maps. Incorporating probabilistic values for both TractSeg and JHU atlas (weighted mean values) aimed to minimised the effect of partial volume effects by marginalising the effect of voxel at ROI boundaries. To compute binarised ROI volumes and weighted mean intensities from the TractSeg probabilistic parcellation a threshold of 0.5 was applied. This was chosen, as it considered voxels to be part of a region when the particular region class was predicted with a 50% certainty. To understand the impact of different scanning protocols on the diffusion MRI derived features, the same methods were also applied to the *Multi-Acquisition* database. This included 16 healthy subjects who underwent DWI imaging on the same scanner, but with five different protocols, all varying in number and strength of b-values.

Multi-Centre Differences in DTI Acquisition The previous experiments were conducted on strongly regulated databases that included the same subjects. These were either scanned multiple times under the same protocol, but on different scanners (*Scan-Rescan*), or imaged with varying acquisition parameters, but on the same scanner (*Multi-Acquisition*). Moreover, MR scans were all collected at the same imaging site, minimising the bias introduced by different operators or imaging policies. However, for multi-centre studies different scanners and unharmonised MRI protocols may be involved. To examine the variability of diffusion MRI metrics across imaging sites, the previous methods for DWI parcellations were applied to the *Multi-Centre* database (a subset of the healthy controls from the CENTER-TBI imaging database).

After pre-selection of the sites (see above), the QC metrics from the diffusion pipeline were

[6] A threshold of 50% eliminated some of the 20 ROIs, both in DWI and atlas space.

analysed to identify and exclude any outliers due to image corruption. The PIS ratio ranged between 0.24 and 0.66, and no gross outliers were found. Nonetheless, both scans with the highest PIS QC ratio (0.66) and with the highest number of voxels with PIS (2343) were checked visually and deemed to appear normal.[7] Furthermore, both scans with the lowest (43.0) and highest (270.2) SNR, were inspected and found without any obvious artefacts. Brain mask QC ratio ranged between 92.5 % and 105.1%, which was accepted as sufficient. One subject from centre B with the highest average total head motion QC metric (1.06) was examined visually and strong noise artefacts were detected. Hence this scan and the corresponding rescan were excluded. The scan with the next highest average total head motion (0.9863) seemed unaffected. After exclusion of the abnormal scan and rescan already mentioned, none of the other head motion QC metrics led to noticeably different scans and no obtrusive artefacts were found. The NCC values between FA maps and T1w images ranged from 0.43 to 0.69. Apart from susceptibility artefacts, no aberrant deformations were found on the scan with the lowest NCC.

An overview of the final curated dataset is provided in Table 4.1, listing the scanner vendor and model, the number of diffusion sensitised volumes (#Dir.) as well as age and sex. Furthermore, it indicates whether susceptibility distortions were corrected with an extra b_0 volume of opposite phase encoding direction. Additionally, the NCC between the JHU FA map and the subjects FA maps, spatially normalised to JHU space, is shown (NCC_{jhu}, not to be confused with the NCC between coregistered FA maps and T1w images).

After this curation, one scan of both a female and male subject from each scanner was chosen that matched the mean age across all centres (40.6 years) the closest. This pre-selection resulted in 18 subjects with an average age (39.8 years) similar to that of all control subjects, but with a slightly narrower age range. Those were then used to compute a study specific template via tensor registration using DTI-TK [280, 282]. Eventually, the final DTI-TK average FA map was parcellated with the JHU atlas as previously described. All scans (138 from 69 subjects including the ones used to create the template) were then spatially normalised to this template.

Before analysing variance in DTI metrics, the curated dataset was examined for inter-centre compatibility. Therefore, the different QC metrics from the diffusion pipeline were compared across centres. Subsequently the local variability of FA and MD metrics were studied both in native as well as template space. These experiments aimed to investigate the differences present within centres, across same scanner vendor located at different sites, as well as the overall variation in the complete dataset.

[7] Reminder: A high PIS ratio indicated that more voxels with PIS remained after artefact correction. Hence, high values are undesirable.

Table 4.1: Overview of Included Data of CENTER-TBI Controls. Age (in years) and NCC_{jhu} displayed as mean±std.

Centre	Vendor	Model	#Dir.	b_0	Age	Sex	NCC_{jhu}
A	Siemens	Trio	30	no	41.7±15.3	1F/5M	0.758±0.021
B	Siemens	Skyra	30	yes	41.8±11.5	3F/5M	0.776±0.029
C1	GE	MR750w*	32	no	38.0±11.1	4F/4M	0.726±0.016
C2	GE	MR750*	32	no	43.3±11.2	4F/5M	0.754±0.019
D	Siemens	Prisma	30	no	41.6±13.1	3F/6M	0.775±0.017
F	GE	MR750*	32	no	45.1±13.0	3F/6M	0.738±0.018
G	Philips	Ingenia	32	no	31.0±6.2	2F/4M	0.759±0.005
H	Philips	Ingenia	32	yes	39.1±15.3	3F/5M	0.777±0.015
L	Philips	Achieva	32	no	40.2±14.3	5F/1M	0.758±0.010
				Total	40.6±12.4	28F/41M	0.758±0.025
				DTI-TK Template	39.8±9.6	9F/9M	0.974

*GE Discovery MR750 & MR750w models. F=female, M=male

4.2.3 Evaluation Metrics

A common metric to estimate the deviation across scans is the CV, which is defined as the ratio of standard deviation σ and mean μ of a distribution. For ROI analysis there are different approaches over which distribution the CV can be calculated [256]. Here, for images in native space, first the mean intensities in each ROI for each scan were computed, and then the CV was calculated on basis of those means:

$$CV^r_{mean} = \frac{\sigma(\bar{X}_r)}{\mu(\bar{X}_r)} \times 100\% \tag{4.1}$$

where $\bar{X}_r$ is the collection of means of a specific ROI r for all images to be investigated. Analogously, the CV for volumetric measurements was computed by dividing the standard deviation of the volume by the mean volume of the examined cohort (denoted as CV^r_{vols} for region r).

With spatial normalised scans, the CV maps were derived on a voxel wise level, before averaging those variation scores within each ROI:

$$CV^r_{voxel} = \frac{1}{N}\sum_{i=0}^{N}\left[\frac{\sigma(x^r_i)}{\mu(x^r_i)}\right] \times 100\% \tag{4.2}$$

with x^r_i representing the voxel intensities at a specific location i within a ROI r with N number of voxels.

Besides this, the statistical difference of ROI metrics (i.e. region volumes or mean intensities of FA and MD within ROIs) was estimated with an *analysis of variances for repeated measurements* (rm-ANOVA). This was chosen under the assumption of equal variances. To account for multiple comparisons, p-values were adjusted with FDR correction and the more stringent Bonferroni correction. For this the modules AnovaRM and multipletests from the python library statsmodels were used. If statistical differences were detected with rm-ANOVA, a post-hoc paired t-test was applied and likewise controlled for FDR. Corrected p-values for post-hoc tests are marked with *hoc* (p_{fdr}^{hoc}) to avoid confusion with p-values from rm-ANOVA (p_{fdr}). T-tests were conducted with ttest_rel from python library scipy.

For aligned atlases in template space, the overlap between regions was measured by calculating the Dice score with f1_score from the scikit-sklearn python library. The Dice score is generally defined as the size of the intersection of two sets A and B over the sum of sizes of both individual sets:

$$\frac{2\,|A \cap B|}{|A| + |B|} \tag{4.3}$$

4.3 Results

4.3.1 Reproducibility of Anatomical Brain Parcellation

Comparing total brain volumes and four MALP-EM tissue compartments (i.e. ventricles, non-cortical GM, cortical GM and WM) within and across scanners showed very similar average volumes on both scanners. Testing volume differences with rm-ANOVA and FDR correction, identified total brain ($p_{fdr} < 0.001$) and WM ($p_{fdr} = 0.0248$) volumes to be significantly different (cortical GM volume differences were close to significance: $p_{fdr} = 0.0615$). The post-hoc paired t-tests revealed only a statistically relevant difference of total brain volume between (all: $p_{fdr}^{hoc} < 0.001$), but not within scanners (Prisma: $p_{fdr}^{hoc} = 0.1707$, Trio: $p_{fdr}^{hoc} = 0.8460$). The total brain volume discrepancy across scanners was on average 10.5 cm^3, which corresponds to approximately 11 voxels or a 1% volume difference, which can be considered marginal. White matter volumes were different between the rescan group on Prisma (Prisma #2) and both scan groups on Trio (Trio #1: $p^{hoc} = 0.0262$, Trio #2: $p^{hoc} = 0.0191$), however, this was not statistically significant after FDR correction (for both inter-scanner: $p_{fdr}^{hoc} = 0.0785$). The WM volume difference for this data corresponded to approximately 8 voxels, or a discrepancy of 1.6%. All volumes for tissue compartments and total brain are summarised in Table 4.2 .

In a rm-ANOVA 47 ROIs had a p-value<0.05, however, after correction for multiple com-

Table 4.2: MALP-EM Volumes for Tissue Compartments. Volumes displayed as mean ± std in cm^3

Scan	Total Brain	Ventricles	Non-cGM	cGM	WM
Prisma #1	1241.2 ± 140.7	20.2 ± 6.4	179.9 ± 14.5	564.9 ± 63.8	474.7 ± 64.5
Prisma #2	1242.8 ± 140.9	20.2 ± 6.4	178.1 ± 12.2	566.0 ± 61.5	477.1 ± 67.4
Trio #1	1231.4 ± 141.8	20.2 ± 6.4	179.0 ± 12.5	562.0 ± 58.9	468.7 ± 72.0
Trio #2	1231.6 ± 142.0	20.2 ± 6.3	178.1 ± 14.2	561.9 ± 59.8	469.9 ± 70.5

parison with FDR only 34 ROIs were significantly different. This number further decreased to 13 with more stringent Bonferroni correction.

Volume discrepancies were found mostly across scanners, and only the brainstem was significant different between both Prisma scanners ($p_{fdr}^{hoc} = 0.0132$). Regions in both hemispheres with volumetric differences across scanners included the cerebellar WM ($p_{fdr} \leq 0.0158$), the palladium ($p_{fdr} \leq 0.0298$), the thalamus proper ($p_{fdr} \leq 0.0025$), the ventricle DC ($p_{fdr} \leq 0.0155$), the superior temporal ($p_{fdr} \leq 0.0011$) and lateral orbital gyrus ($p_{fdr} \leq 0.0213$), the central operculums ($p_{fdr} \leq 0.0074$) as well as the frontal ($p_{fdr} \leq 0.00337$) and occipital poles ($p_{fdr} \leq 0.0001$). Apart from left and right cerebral WM compartments being the largest ROIs ($\approx 218\,cm^3$) in the MALP-EM atlas, all other mentioned ROIs were neither particularly large or small. Figure 4.1 shows four ROI examples for both equal and deviating volumes across scanners. There was no systematic bias of volumes being consistently larger on Prisma or Trio.

When comparing relative ROI segmentations, the number of ROIs with deviating volumes on different scans could be reduced (no correction: 40) and was approximately halved when considering multiple comparison correction (FDR: 17, Bonferroni: 8).

Instead of computing relative volumes post-parcellation, image intensities were harmonised via NDFlow before applying MALP-EM. Matching intensities from Prisma scanners to the Trio intensity space reduced the volume discrepancy to 17 ROIs with significantly different volumes across scanners. After multiple comparison correction with FDR or Bonferroni this was decreased to seven ROIs or one region, respectively. Among the seven ROIs identified with different volumes across centres were the left and right WM of the cerebellum (ROI #12 & #13), the left putamen (ROI #28), the left thalamus proper (ROI #28), the right ventral DC (ROI #31), the right lateral orbital gyrus (ROI #71) and the right occipital pole (ROI #91). All ROIs that were different after intensity matching were also different without this pre-processing step, implying the intensity matching does not introduce any volume deviation itself. The p-values of the post-hoc t-test after FDR correction are listed in Table 4.3. Visual inspection of ROIs with and without intensity matching showed only

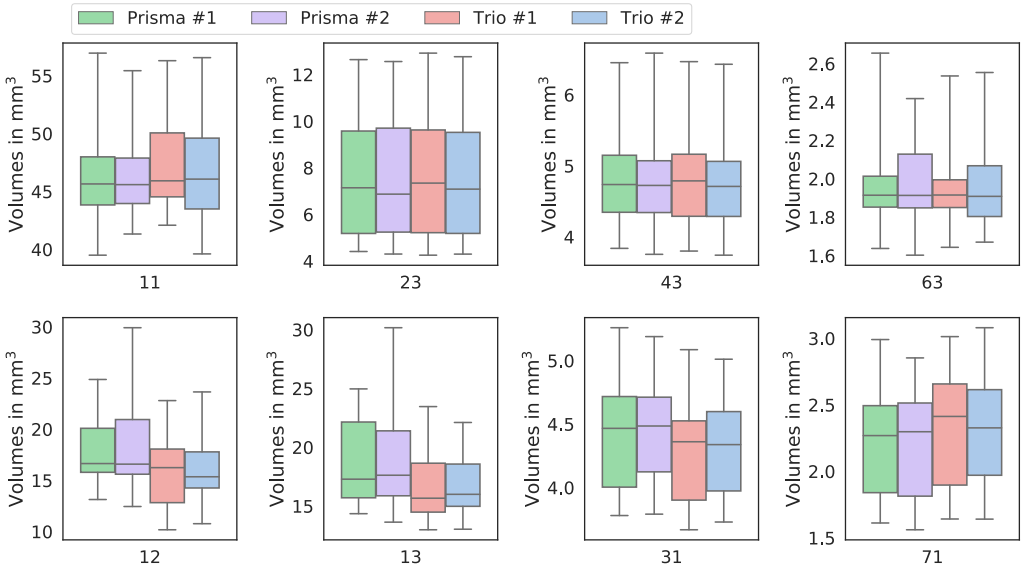


Figure 4.1: Distribution of Volumes from MALP-EM Regions Deviating Across Scanners. **Top row:** ROIs which were segmented cohesive on all four T1w scans and were found to be statistically indifferent (ROI #11, 23, 43, 63). **Bottom row:** ROIs with volumes that were statistically different across, but not within scanner (ROI #12, 13, 31, 71). Deviations between both scanners were minimal. No systematic tendencies for over- or undersegmentation was detected on either of the scanners. **ROIs:** 11: Left cerebellum exterior, 23: Right lateral ventricle, 43: Right anterior insula, 63: Right gyrus rectus, 12: Right cerebellum WM, 13: Left cerebellum WM, 31: Right ventral DC, 71: Right lateral orbital gyrus. **Boxplots:** In each subplot from left to right: Prisma #1, Prisma #2, Trio #1, Trio #2.

minimally deviating volumes. Despite the positive effect of matching intensities to the study-independent Cam-CAN database prior to the parcellation, regional volume differences were still present. The number of different MALP-EM ROIs across scanners was almost identical (no correction 46, FDR: 31, Bonferroni: 13) as when no intensity harmonisation was applied.

Lastly, the analysis of MALP-EM volumes derived from scans previously coregistered showed a slightly reduced number of ROIs with segmentation discrepancies. Forty-one regions were highlighted as significantly different, whereas 24 or 11 remained after FDR and Bonferroni correction, respectively. Although the same mask was provided initially, the total brain volume was significantly different between Trio and Prisma scanners (inter-scanner comparisons: $p_{fdr}^{hoc} \leq 0.0007$), but not within scanner (Prisma: $p_{fdr}^{hoc} = 0.0867$, Trio: $p_{fdr}^{hoc} = 0.8450$). Apart from the lateral orbital gyrus, same regions[8] were found to have different volumes as when computing MALP-EM in native space. Intra-scanner differences were found on only for Prisma scans for the brainstem ($p_{fdr}^{hoc} \leq 0.0182$) the right pallidum ($p_{fdr}^{hoc} \leq 0.0485$) and

[8]cerebellar WM, palladium, the thalamus proper, the ventricle DC, the superior temporal gyrus, the central operculums, frontal and occipital poles

Table 4.3: P-Values After FDR Correction for Comparison of MALP-EM ROI Volumes Within and Across Scanners After Intensity Matching. P-values of post-hoc t-test $p^{hoc}_{fdr} < 0.05$ printed in bold.

		Intra-Scanner		Inter-Scanner			
ROI	rm-ANOVA	P1-P2	T1-T2	P1-T1	P1-T1	P2-T1	P2-T2
12	0.0190	0.5301	0.3487	**0.0112**	**0.0372**	**0.0131**	**0.0333**
13	0.0190	0.8589	0.5438	**0.0114**	**0.0114**	**0.0402**	**0.0440**
28	0.0259	0.5465	0.6011	**0.0010**	**0.0002**	0.1638	0.1061
30	0.0190	0.1599	0.7383	**0.0074**	**0.0002**	0.1599	0.1063
31	0.0190	0.2244	0.2244	**0.0137**	**0.0197**	**0.0197**	**0.0363**
71	0.0083	0.0899	0.6355	**0.0112**	**0.0016**	**0.0091**	**0.0066**
91	0.0223	0.0978	0.7358	**0.0034**	**0.0042**	0.2687	0.3924

ROIs: 12: right cerebellum WM, 13: left cerebellum WM, 28: left putamen, 30: left thalamus proper, 31: right ventral DC, 60: left frontal pole, 71: right lateral orbital gyrus, 91: right occipital pole. **Scans:** P1: Prisma scan #1, P2: Prisma scan #2, T1: Trio scan #1, T2: Trio scan #2.

the left middle frontal gyrus ($p^{hoc}_{fdr} \leq 0.0446$).

Since T1w images were previously registered to one and another, the MALP-EM atlases were also aligned, allowing computation of the overlap for each region. The Dice scores for all regions were on average marginally higher on Prisma ($Dice = 0.930 \pm 0.034$) than on Trio ($Dice = 0.925 \pm 0.037$) scanners. Overlap of regions were overall lower when comparing images across scanners (Prisma #1 vs Trio #1: $Dice = 0.896 \pm 0.043$ and Prisma #2 vs Trio #2: $Dice = 0.893 \pm 0.046$). This finding is in agreement with the previous observations of more significant volume deviations for inter-scanner than for intra-scanner comparisons.

4.3.2 Variation of White Matter Region Segmentation

TractSeg. The rm-ANOVA on the TractSeg volume revealed significant differences ($p < 0.05$) for 23 out of 72 fibre tracts. Correcting for multiple comparisons reduced this number to 13 (FDR) and nine (Bonferroni). The p-values for the 13 ROIs with statistical differences between the four groups (scan and rescan on both scanners) are listed in Table 4.4. A post-hoc t-test with FDR correction revealed only significant p-values for ROI volume differences across scanners. Intra-scanner volume measurements were comparable to one another. Particularly important are the eight ROIs which showed inter-scanner differences for all four scan-rescan combinations, such as for example the right SLF_I (ROI #36) or the CC (ROI #45). Volumes that were observed to be different only for one scan pair, such as the genu of

CC (ROI #6) or the left thalamic premotor cortex (ROI #48), potentially highlight single outliers. Noteworthy, the rostral body of CC (ROI #7) as well as the left inferior cerebellar peduncle (ROI #22) were found to have different inter-scan volumes (both rm-ANOVA and uncorrected post-hoc t-test, p-values < 0.05), however, FDR correction eliminated this statistical difference.

Table 4.4: P-Values After FDR Correction for Comparison of TractSeg ROI Volumes Within and Across Scanners. P-values of post-hoc t-test <0.05 printed in bold.

		Intra-Scanner		Inter-Scanner			
ROI	rm-ANOVA	P1-P2	T1-T2	P1-T1	P1-T1	P2-T1	P2-T2
0	0.0023	0.3460	0.4776	**0.0210**	**0.0210**	**0.0210**	**0.0210**
6	0.0365	0.9091	0.2230	0.1042	**0.0266**	0.1042	0.1042
7	0.0274	0.8333	0.8333	0.0736	0.0717	0.0717	0.0717
22	0.0365	0.6104	0.9630	0.0759	0.0759	0.0759	0.1004
25	0.0005	0.1192	0.3390	**0.0090**	**0.0107**	**0.0032**	**0.0032**
30	0.0040	0.2038	0.5703	**0.0339**	**0.0339**	**0.0223**	**0.0223**
31	0.0038	0.4538	0.7709	**0.0118**	**0.0118**	0.0680	0.0611
32	0.0008	0.9094	0.6386	**0.0092**	**0.0105**	**0.0092**	**0.0105**
36	0.0040	0.2865	0.2862	**0.0388**	**0.0388**	**0.0388**	**0.0388**
45	0.0009	0.9420	0.1751	**0.0058**	**0.0094**	**0.0167**	**0.0256**
48	0.0453	0.4980	0.6083	0.1105	**0.0978**	0.1105	0.1105
57	0.0023	0.2598	0.5339	**0.0162**	**0.0162**	**0.0099**	**0.0099**
71	0.0040	0.7229	0.8952	**0.0025**	**0.0205**	**0.0205**	**0.0242**

ROIs: 0: left arcuate fascicle, 6: genu, 7: rostral body of corpus callosum, 22: left inferior cerebellar peduncle, 25: right IFO, 30: right optic radiation, 31: left parieto-occipital pontine, 32: right parieto-occipital pontine, 36: right SLF_I, 45: CC, 48: left thalamo-premotor tract, 57: right thalamo-occipital tract, 71: right striato-occipital tract. **Scans:** P1: Prisma scan #1, P2: Prisma scan #2, T1: Trio scan #1, T2: Trio scan #2.

Figure 4.2 shows the volume distributions of four ROIs without discrepancies (top row) and four of the ROIs that were found statistically different (bottom row). No systematic bias of consistently smaller volumes on one or the other scanner was found. Some ROIs had a larger volume on Trio, such as the left arcuate fascicle (ROI #0) or the CC (ROI #45). Other ROIs, such as the left IFO (ROI #25) or the right SLF_I (ROI #36) were larger on Prisma scans.

JHU. Parcellating the brain scans via projection of the JHU-atlas ROIs did not show any obvious volumetric differences across the four scan session. The rm-ANOVA highlighted six ROIs to be potentially different ($p < 0.05$), however, after multiple comparison correction with either Bonferroni or FDR none of the ROIs examined showed a significantly different volume.

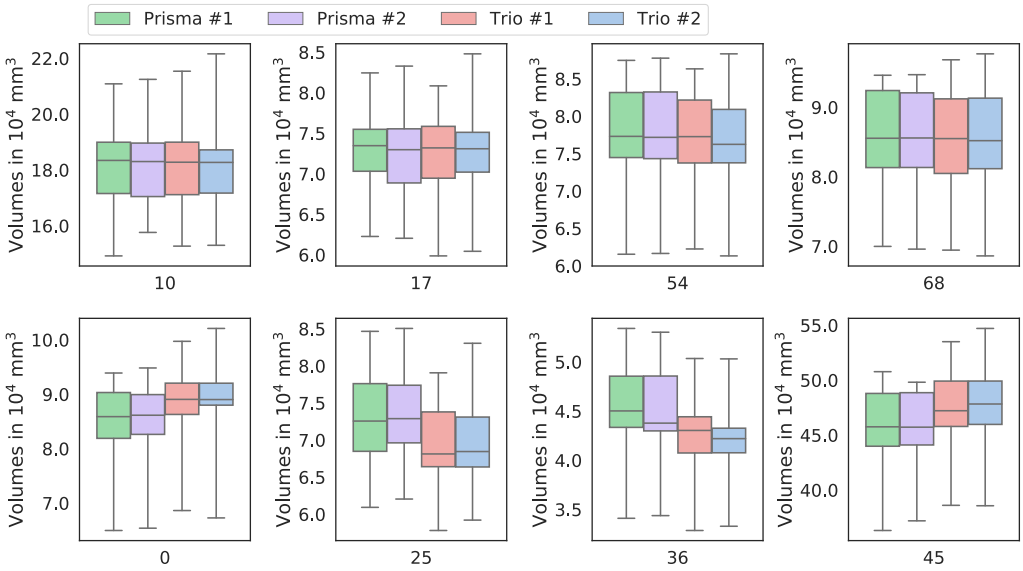


Figure 4.2: Distribution of Volumes from TractSeg Regions Deviating Across Centres. **Top row:** Four examples of TractSeg ROIs with highly reproducible volumes. **Bottom row:** Some regions showed deviating volumes when comparing across scanners, while remaining similar on the scanner. **ROIs:** 10: Isthmus, 17: Right middle longitudinal fascicle, 54: Left thalamo-parietal tract, 68: Left striato-parietal tract, 0: left arcuate fascicle 25: right IFO fascicle, 36: right SLF_I, 45: CC. **Boxplots:** In each subplot from left to right: Prisma #1, Prisma #2, Trio #1, Trio #2. Whiskers show the full range of the distribution.

4.3.3 Inter-Scanner Differences of DTI Metrics

After assessing the volume differences, local variance in DTI parameter maps was measured by comparing weighted means of FA and MD within TractSeg and JHU ROIs.

TractSeg FA. Out of the 72 fibre tracts, 25 ROIs were flagged as having a significantly different FA ($p < 0.05$). From those, mean FA values of 14 ROIs remained statistically different after FDR correction ($p_{fdr} < 0.05$). This was further reduced to five ROIs when correction for multiple comparison with Bonferroni. Figure 4.3 shows four selected ROIs which were detected as different by the rm-ANOVA analysis alongside ROIs with comparable inter-scanner mean FA values. Regions with significantly different FA values between Prisma and Trio scans showed no consistent bias on one of both scanners. Lower FA mean values were found for Prisma than for Trio in regions such as for example the right superior thalamic radiation (ROI #41: $p_{fdr} < 0.001$, Prisma vs Trio: $p^{hoc}_{fdr} < 0.002$). A similar trend was observed in regions such as the left CST (ROI #14: $p_{fdr} = 0.0060$, Prisma #1 vs Trio: $p^{hoc}_{fdr} < 0.0076$) or the left striato-postcentral tract (ROI #66: $p_{fdr} = 0.0163$, Prisma #1 vs Trio: $p^{hoc}_{fdr} < 0.0274$). For both regions no statistical significance between Prisma #2 and Trio scans was measured (ROI #14: $p^{hoc}_{fdr} > 0.0610$, ROI #66: $p^{hoc}_{fdr} > 0.1190$). In contrast, the left striato-fronto-orbital tract (ROI #58: $p_{fdr} = 0.0040$) exhibited lower mean FA values on Trio scans than Prisma scans (all post-hoc paired t-tests between Prisma and Trio scans $p^{hoc}_{fdr} < 0.05$). No intra-scanner differences were detected for the four mentioned regions. A detailed list of ROIs with deviating mean FA values and p-values is provided in the Appendix (Table A.1).
Calculating the CV of the mean FA within ROIs revealed that the inter-scanner variances ($CV_{mean} = 3.2\% \pm 1.2\%$) was comparable to the ones found on Prisma ($CV_{mean} = 3.0\% \pm 1.2\%$) and Trio ($CV_{mean} = 3.1\% \pm 1.4\%$). On Prisma scans all CV scores for individual ROIs lay below 9%. For Trio scans only the right fornix ($CV^{21}_{mean} = 12.5\%$) was above the 10% threshold. This was also reflected when computing the inter-scanner CV scores ($CV^{21}_{mean} = 10.6\%$).

TractSeg MD. More ROIs with significantly different mean values were found for MD maps (41). Almost half of the segmented fibre tracts (30) showed different mean MD values between the scanners after FDR correction ($p_{fdr} < 0.05$). With Bonferroni correction nine ROIs remained statistically different. Similar to FA, there was no consistent bias and higher average MD values were found in ROIs both on Prisma and Trio. Four representative ROIs with no inter-scanner and with obvious differences are presented in Figure 4.4. So

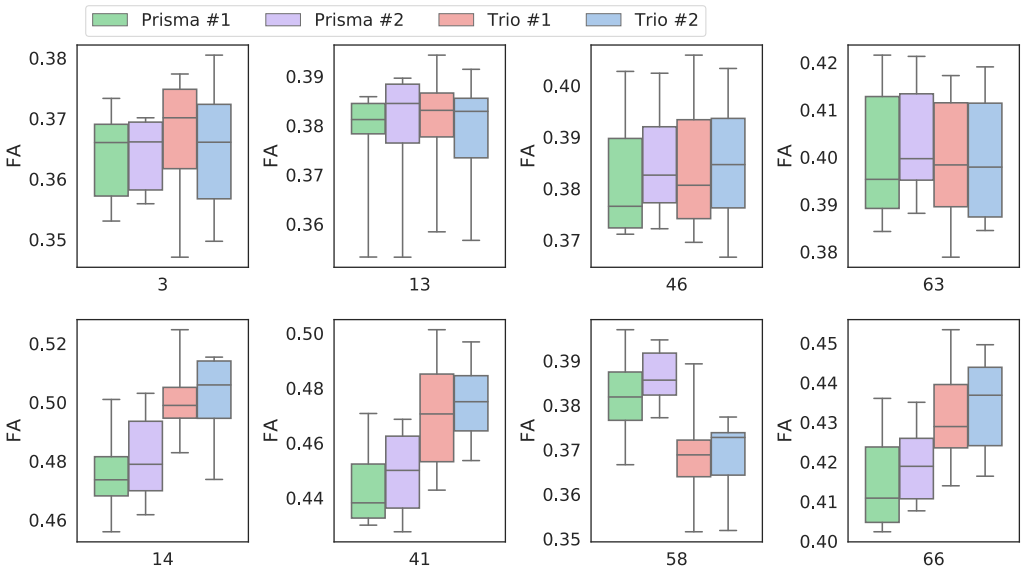


Figure 4.3: Differences of Average FA Intensities within Selected TractSeg ROIs. **Top row:** ROIs where no statistical difference was found. Intra- and inter-scanner variation was comparable. **Bottom row:** ROIs for which significant differences were observed. Scanner-specific biases are inconsistent, so that both elevated and decreased means could be found on Trio compared to Prisma. **ROIs:** 3: Right anterior thalamic radiation, 13: Right cinculum, 46: Left thalamo-prefrontal tract, 63: Right striato-premotor tract, 14: Left CST, 41: Right superior thalamic radiation 58: left striato-fronto-orbital tract, 66: Left striato-postcentral tract. **Boxplots:** In each subplot from left to right: Prisma #1, Prisma #2, Trio #1, Trio #2. Whiskers show the full range of the distribution.

for example the middle and superior cerebellar peduncles (ROI #28: $p_{fdr} = 0.0004$; ROI #33: $p_{fdr} = 0.0016$) showed higher MD values on Prisma than Trio. Both, however, had lower significance levels for the Prisma rescan (Prisma #2) in comparison to Trio scans, than the initial Prisma scan (Prisma #1). Contrary, higher MD values were found on Trio scans for the left thalamo- and striato-postcentral tracts (ROI #51: $p_{fdr} = 0.0103$; ROI #67: $p_{fdr} = 0.0004$). Again, all four combinations of inter-scanner comparisons showed differently pronounced deviations. All 30 ROIs had comparable MD values for scans and rescans on both Prisma and Trio scans. An overview of p-values is provided in the Appendix (Table A.2).

Comparing the CV of mean MD values within ROIs revealed a higher variability for Trio scans ($CV_{mean} = 2.4\%\pm1.2\%$) than for Prisma scans ($CV_{mean} = 2.0\%\pm1.2\%$). This was also propagated, but not further elevated, for inter-scanner comparison ($CV_{mean} = 2.4\%\pm1.2\%$). All CV scores fell below the 10% mark except for the MD values in the right fornix ($CV^{21}_{mean} = 10.1\%$) on Trio scans. Overall CV was lower for MD than for FA metrics.

Six TractSeg ROIs (#14, #18, #28, #33, #41 & #45) showed significant inter-scanner

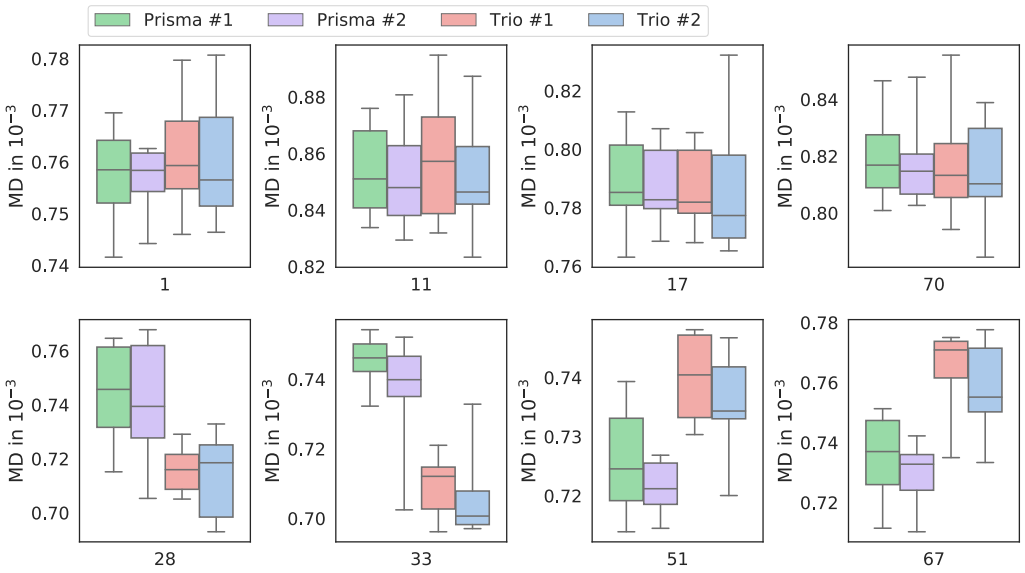


Figure 4.4: Differences of Average MD Intensities within Selected TractSeg ROIs. **Top row:** ROIs which appeared to have a comparable distribution between all 4 scans. **Bottom row:** ROIs for which significant deviations were found. Means within ROIs were increased as well as reduced for Trio scans relatively to Prisma. **ROIs:** 1: Right arcuate fascicle, 11: Splenium, 17: Left middle longitudinal fascicle, 70: Left striato-occipital tract, 28: Middle cerebellar peduncle, 33: Superior cerebellar peduncle, 51: Left thalamo-postcentral tract, 67: Left striato-postcentral tract. **Boxplots:** In each subplot from left to right: Prisma #1, Prisma #2, Trio #1, Trio #2.

differences for both FA and MD. Among those only the CC (ROI #45) was previously reported to show a different volume across scanners.

JHU FA. Comparing mean DTI metrics in the 20 JHU ROIs via rm-ANOVA and FDR correction on the same dataset, revealed mean FA differences for seven ROIs (with and without FDR correction). This was reduced to four ROIs after Bonferroni correction. Similar to TractSeg, the FA inter-scanner variability ($CV_{mean} = 4.6\% \pm 2.1\%$) was comparable to that within scanners. However, CV scores were on average slightly lower for Trio ($CV_{mean} = 4.2\% \pm 2.0\%$) than for Prisma ($CV_{mean} = 4.7\% \pm 2.4\%$) or both datasets combined. Thirteen out of 20 ROIs showed lower individual CV scores on Trio than on Prisma. Only the left CG had a high variability above 10% for both Prisma scans and inter-scanner comparison ($CV^4_{mean} = 11.7\%$ and $CV^4_{mean} = 10.1\%$, respectively). Average variability scores for mean FA values calculated within JHU ROIs were higher than for TractSeg ROIs.

JHU MD. Analysis of mean MD differences (rm-ANOVA) identified ten ROIs after FDR correction as significantly different across scanners (eleven ROIs without correction and

seven ROIs after Bonferroni correction). Intra-scanner variation of MD mean values in JHU ROIs was practically equal for Prisma and Trio ($CV_{mean} = 2.4\% \pm 0.8\%$ and $CV_{mean} = 2.4\% \pm 0.7\%$, respectively), but slightly elevated when fusing datasets from both scanners ($CV_{mean} = 2.8\% \pm 0.8\%$). Coherent with the previous TractSeg findings, the variation found for MD was lower than for FA maps.

Six of the 20 JHU ROIs were different for both FA and MD, including the left and right anterior thalamic radiation (ROI #0: $p_{fdr} < 0.030$; ROI #1: $p_{fdr} < 0.002$), the left CST (ROI #2: $p_{fdr} < 0.001$), the hippocampal cingulum bundle (ROI #6: $p_{fdr} < 0.004$) and the (temporal) left SLF (ROI #14: $p_{fdr} \leq 0.020$; ROI #18: $p_{fdr} < 0.05$). For a complete list of p-values see Appendix Table A.3.

4.3.4 Impact of Acquisition Protocol on Fibre Tract Segmentation

Analysing TractSeg volumes via rm-ANOVA identifies significant differences ($p_{fdr} < 0.05$) for all regions, but the right fornix (ROI #21), across acquisition protocols. TractSeg ROI segmentations were found to be dependent on the number of shells and directions the DWI image was acquired with. Out of 72 TractSeg ROIs, 69 showed volumes that were on DTI 20×3 larger than on DTI 12×5, but simultaneously smaller on DTI 60×1 (DTI 12×5 < DTI 20×3 < DTI 60×1). A strict staircase pattern (Figure 4.5), where protocols with less shells and more directions showed larger volumes (DTI 12×5 < DTI 15×4 < DTI 20×3 < DTI 30×2 < DTI 60×1), was observed for more than 60% (i.e. 45) of the ROIs. Comparing volumes of JHU fibre tract segmentations did not reveal any significant differences after FDR correction of rm-ANOVA results.

4.3.5 Acquisition Protocol Specific Differences in DTI Metrics

TractSeg. Analysis of mean FA and MD values within TractSeg ROIs highlighted all regions as significantly different (rm-ANOVA $p_{fdr} < 0.05$). When comparing mean FA or MD values of DTI 60×1 to any other protocol, significant differences ($p^{hoc}_{fdr} < 0.05$) were found almost for all regions. Discrepancy was lowest for DTI 12×5 and DTI 15×4, with twelve or six regions observed significantly different for FA and MD metrics, respectively. Fewer significant regions were found for protocols that had a more similar angular resolutions (and number of shells) than for those with more deviating acquisition parameters. For example when comparing DTI 15×4 to DTI 30×2 37 ROIs showed significant different mean MD values, but compared to DTI 20×3 only six regions were detected.
Comparing the variation of TractSeg ROIs for data from single protocols against each other

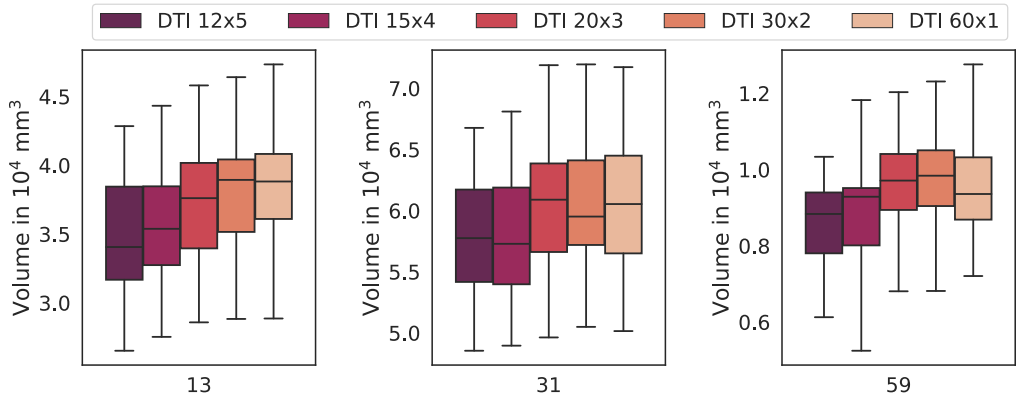


Figure 4.5: Reputability of Tract Volumes for Different DWI Acquisitions. Most ROIs showed the same staircase-pattern for volumes as found for the right cingulum (ROI #13). Larger ROI volumes were segmented on scans acquired with more directions and less shells. Nonetheless, there were a few ROIs which disrupted this pattern such the here displayed ROI #31 and #59. **ROIs: 13:** Right cinculum, 31: Left parieto-occipital pontine tract, 59: Right striato-fronto-orbital tract. **Boxplots:** In each subplot from left to right: Protocols with different number of directions per shells [dir×shell]. DTI 12×5, DTI 15×4, DTI 20×3, DTI 30×2, DTI 60×1.

showed similar levels of $CV_{mean} \approx 5.0\%$ for FA. Only a significant difference was found between data from protocol DTI 12×5 and DTI 30×2 ($p^{hoc}_{fdr} = 0.0214$). The TractSeg MD ROI variation was similar for all acquisitions ($CV_{mean} \approx 4.1\%$) except for DTI 60×1, which showed a lower regional MD variance ($CV_{mean} = 3.8\%$) than any other protocol ($p^{hoc}_{fdr} < 0.0310$).

Comparing TractSeg's CV for volumes (CV_{vols}) against CV_{means} for FA or MD values for all single protocol data (intra-protocol variation showing differences between subjects) and combinations of two protocols (inter-protocol variation showing differences between acquisition parameters) revealed only a weak correlation (Spearman's correlation: r_{FA}=0.4318 and r_{MD}=0.1376). When plotting CV_{means} for FA against CV for TractSeg ROI volumes (CV_{vols}, Figure 4.6 Left) the variation of mean FA values in TractSeg ROI mostly clustering together around the 5% mark. However, with increasing variation in ROI volumes ($CV_{vols} > 30\%$) the variations for regional FA mean values also increased to variations of 8% to 12%. Analogous for MD (Figure 4.6 Right), the variations for different TractSeg ROIs clustered around 4% and outliers with variations between 10% to 19% were detected mostly for highly deviating volumes ($CV_{vols} > 30\%$). For both FA and MD this increased variance was not exclusive to *Inter-Protocol* observations, but also to data from one single acquisition scheme. Identifying the ROIs that showed the highest CV of volumes within data from the same protocol flagged the anterior commissure ($CV^4_{means} = 27.1\%$) as well as the left and right fornix ($CV^{20}_{means} = 28.0\%$, $CV^{21}_{means} = 38.1\%$). All three regions have on

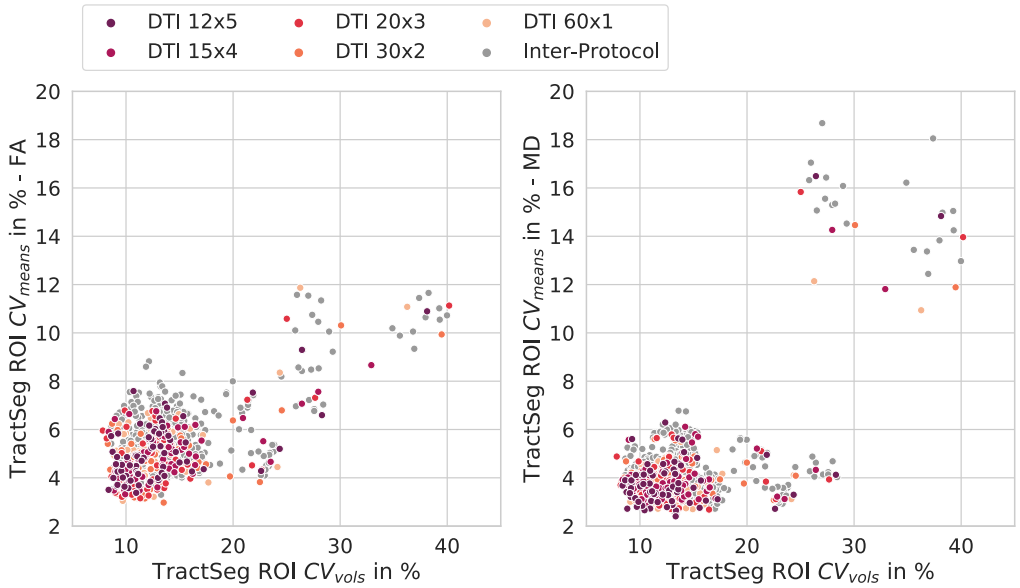


Figure 4.6: Relationship Between CV of DTI Metrics and Volumes of TractSeg Parcellation. **Left:** CV values for FA are clustered between 3%-8% for regions with lower volumetric variation. TractSeg ROIs with volume deviations above 25% also displayed increased discrepancy in FA values (8%-12%). **Right:** CV values for MD mostly score between 2%-6%. Similarly to FA, difference in regional MD values increased (12%-19%) with elevated volume variation. Each dot represents one ROI comparison.

average the lowest volumes (4874, 905 and 794 mm^3, respectively). Indeed excluding these three volumes also eliminates all outliers detected on Figure 4.6.

JHU. Mean values in all JHU regions were identified to be significantly different across different acquisition protocols. On average mean FA values were found to be lower for scans with higher angular resolution per shell (i.e. DTI 30×2 and DTI 60×1), which held true particularly for DTI 60×1. While data with three or more shells (i.e. DTI 12×5, DTI 15×4 and DTI 20×3) showed similar mean FA distributions in JHU tracts, strong significant differences were observed when compared to DTI 30×2 and DTI 60×1 ($p^{hoc}_{fdr} < 0.001$). The same effect was observed for MD mean values.

The average CV within JHU ROIs on FA maps was highest for DTI 60×1 ($CV_{mean} = 5.5\%$) and was significantly ($p^{hoc}_{fdr} < 0.05$) larger than any other protocol CV_{mean} (ranging from 4.9-5.1%). Average CV_{mean} was overall lower for MD than for FA, ranging from 3.9-4.3%. For MD CV_{mean} in JHU ROIs, a significant difference ($p^{hoc}_{fdr} = 0.05$) was only observed between DTI 12×5 and DTI 15×4. Figure 4.7 displays the CV_{mean} distributions of FA and MD means in JHU ROIs for data from single protocols and any other combination. When pooling data from any arbitrary protocol together with data acquired under the DTI 60×1

scheme, the CV_{mean} for both FA and MD increased. This further underlines the differences of DTI metrics of this protocol compared to any of the other four acquisition schemes.

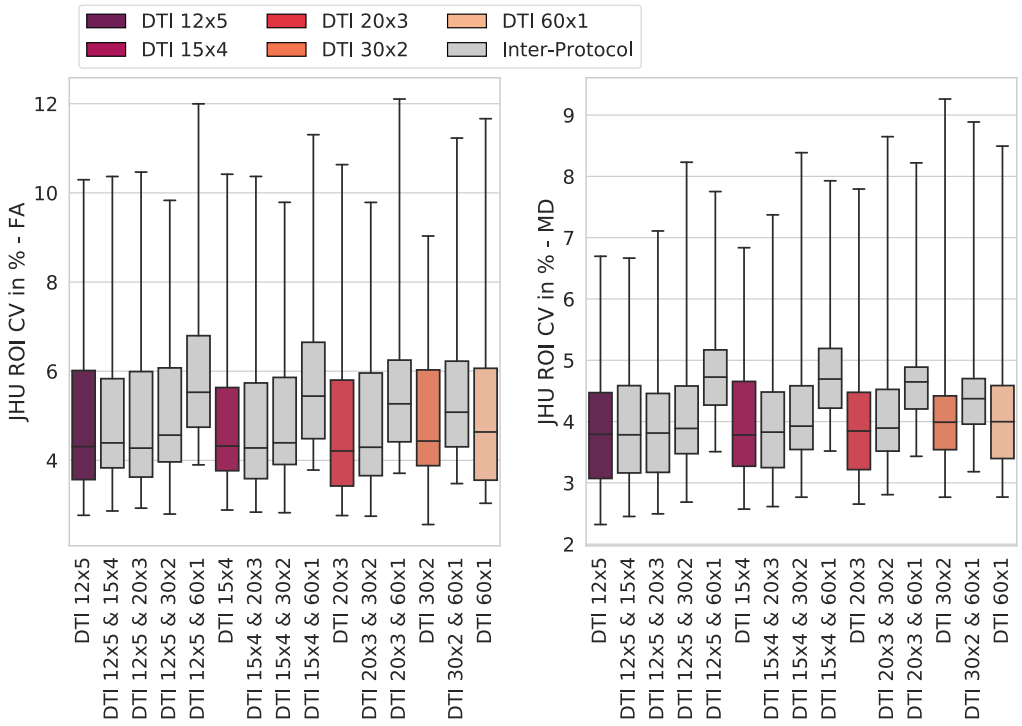


Figure 4.7: Distribution of Regional Coefficients of Variances Across Different Protocols. The intra-protocol CV for all JHU volumes are represented as colour coded boxplots, all combinations of data from different DTI protocols are displayed as grey boxplots. On average variation was found to be lower for MD than fro FA. Overall CV values for both FA and MD are slightly elevated for data acquired with 60 directions and one shell (DTI 60×1), however, data from single protocols showed similar variations ranging between 4-5%. Combining datasets across protocol mostly did not elevate the regional variation, excepts when including the single-shell dataset (DTI 60×1). The latter combined with any other single protocol data increased the average CV for FA as well as MD. Whiskers show the full range of the distribution.

For JHU ROIs, FA deviated most for DTI 60×1 while the MD variation was similar across protocols. This is in contrast to TractSeg ROIs, where FA showed similar variation across acquisition schemes, but MD variation was decreased in scans acquired on one shell (60×1). This discrepancy might be caused by underlying algorithmic methods behind both segmentation approaches (JHU: registration based, TractSeg: model prediction based). An overview of the average CV_{mean} for both DTI metrics and atlases is provided in Table 4.5.

Table 4.5: Average CV of DTI Metrics for JHU and TractSeg ROIs

Atlas	Metric	DTI 12x5	DTI 15x4	DTI 20x3	DTI 30x2	DTI 60x1
TractSeg	FA	5.1±1.3	5.0±1.0	5.0±1.4	4.9±1.3	5.0±1.5
	MD	4.0±2.1	4.1±1.7	4.1±2.0	4.1±1.7	**3.8±1.4**
JHU	FA	5.1±2.2	5.0±1.9	4.9±2.2	4.9±1.8	**5.5±2.4**
	MD	3.9±1.1	4.1±1.1	4.1±1.3	4.3±1.4	4.2±1.3

4.3.6 Comparability of Multi-Centre Data Acquisition

Figure 4.8 displays the different distributions of the QC metrics for the nine scanners included. A high PIS ratio is undesired, as it means that a large proportion of voxels with PIS was still present after correction. For all centres, on average at least 40% of the voxels with PIS were corrected ($PISratio < 0.6$). Centres A, B and D displayed lower mean PIS ratios than all other centres. This may be related to the vendor of the MRI scanner, as those three centres collected data on Siemens scanners ($PISratio = 0.360 \pm 0.077$). Philips scanners (G, H, L) showed on average lower PIS ratios for all three scanners ($PISratio = 0.466 \pm 0.055$), than the three GE scanners (C1, C2, F; $PISratio = 0.583 \pm 0.033$). Distributions of PIS ratio were highly comparable for scan and rescans within centres.

Although differences between centres were observable, the SNR on b_0 volumes was overall fairly evenly distributed for most centres with average values ranging between 126.2 and 196.1. The exception was centre A which showed a much lower average SNR ($SNR = 58.1 \pm 9.8$) than any other centre. All centres, including A, showed a reproducible SNR distribution on the rescans.

Different levels of NCC between T1w images and coregistered FA maps were observed across centres. No particular pattern was recognisable, however, centre G had a much lower mean NCC than any other centre ($NCC = 0.464 \pm 0.023$). All coregistered FA maps for centre G were visually inspected through the QC images provided by the diffusion pipeline and no failed coregistration was discovered. Since FA maps seemed adequately coregistered on visual inspection and intensities are inherently scaled between [0,1], lower NCC scores potentially were a consequence of different intensity ranges of the reference T1w images. Indeed, T1w scans scanned at centre G had much higher average intensities within the brain mask (1518.5 ± 337.6) than for example T1w images from centre H (223.5 ± 37.7), for which high NCC values were found ($NCC = 0.651 \pm 0.019$), despite images collected on scanners from the same vendor (Philips).

Similarly, centre G showed low brain mask volume ratios ($93.6\% \pm 0.8\%$), which could indicate an oversegmentation of the brain when backprojecting the T1w mask to DTI space.

Visual inspection confirmed this small discrepancy between both brain masks (T1w mask back projected and MRtrix3 brain mask, see Section 2.4.2 QC #3). Although the backprojected T1w mask did indeed slightly oversegment the brain, the MRtrix3 brain mask tended to cut of part of the brain. Since brain mask volume ratios on H and L showed the 2nd and 3rd lowest average values ($96.4\% \pm 1.5\%$ and $95.0\% \pm 1.8\%$, respectively), this could suggest a systematic bias on Philips scanners.

Less scattered values were seen for all four motion QC metrics, and values were comparable for scans repetitions on individual MRI scanners. However, one exceptions was centre C1 that particularly stood out regarding average and maximum relative head motion (bottom row Figure 4.8). Interestingly, C1 and C2 are both GE scanners at the same imaging site. Therefore, it was expected that operational differences (e.g. different fixation of head) were minimal. Since control subjects were fairly matched in age and sex (Table 4.1), an increased head motion was not expected due to vastly different cohorts. Most prominent difference in acquisition parameter was the TR which was almost twice as long for scans at centre C1 (TR = 15555 ms) than at centre C2 (TR = 8000 ms). Since head motion and shorter TR at centre C2 were both comparable to other imaging sites, the observed increased head motion at centre C1 is likely a direct consequence of the prolonged scan time. The average values for all QC metrics for each centre are summarised in Table 4.6.

4.3.7 Variation of DTI Metrics for Multi-Centre Data

TractSeg & JHU rm-ANOVA. When comparing volume distributions of TractSeg ROIs across centres, a strong undersegmentation of fibre tracts was observed for centres that employed a Philips scanner (G, H, L). An example of two volume distributions is shown in the Appendix (Figure A.1). Concluding that TractSeg did not provide satisfactory brain parcellations for Philips data, the three centres were excluded from any further TractSeg ROI analysis. Comparing volumes of segmented fibre tracts across scanners showed significant differences (rm-ANOVA $p_{fdr} < 0.05$) for most regions (64) of the TractSeg parcellation. This was also reflected for FA (65) and MD (59) metrics across the six scanners. Regions that showed no significant differences for all three metrics (volumes, FA and MD) included the left fornix (ROI #20), the right IFO fascicle (ROI #23) and the right SLF_I (ROI #34). The JHU segmentation appeared to be more robust, with significant volume differences only for the forceps major (ROI #8) as well as the left and right ILF (ROI #12 & #13). Regional mean FA values statistical deviated across all nine scanners for more than half of the JHU regions. Mean MD values were significantly different for all 20 ROIs.

TractSeg CV. Furthermore, the variance within and across scanners was measured via

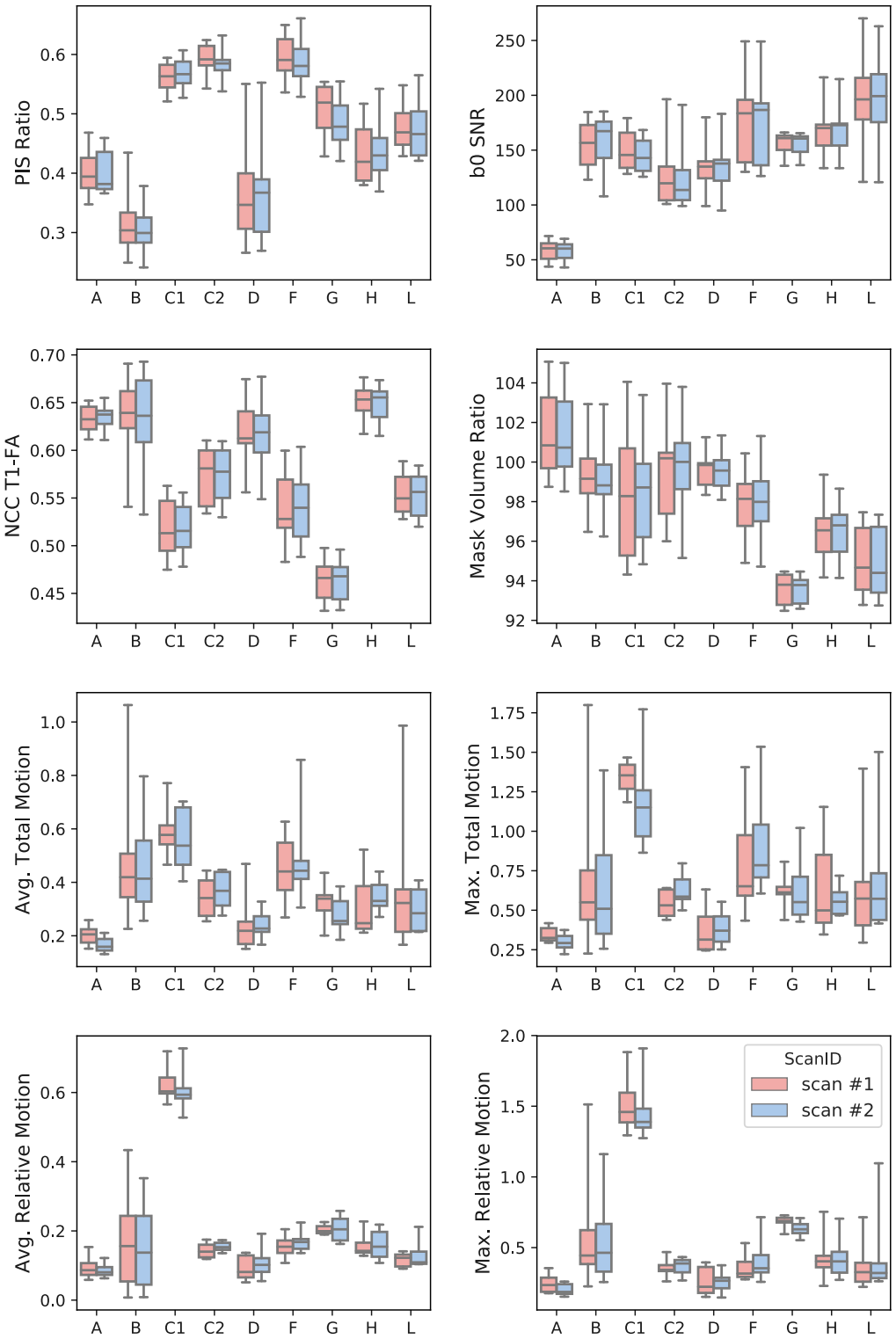


Figure 4.8: Distribution of Quality Control Metrics for Control Subjects from the CENTER-TBI Database

Table 4.6: Quality Controls Metrics Across Scanner for CENTER-TBI Controls. Average values that were noticeably different are printed in bold. All values displayed as mean ± std

Centre	PIS Ratio	b_0 SNR	NCC T1-FA	Mask Volume Ratio in %	Avg. Total Head Motion	Max. Total Head Motion	Avg. Relative Head Motion	Max. Relative Head Motion
A	0.402 ± 0.039	**58.1** ± 9.3	0.634 ± 0.014	101.4 ± 2.3	0.184 ± 0.037	0.322 ± 0.055	0.090 ± 0.026	0.224 ± 0.059
B	**0.312** ± 0.053	156.1 ± 23.5	0.633 ± 0.046	99.4 ± 2.0	0.482 ± 0.222	0.680 ± 0.430	0.164 ± 0.134	0.576 ± 0.350
C1	0.565 ± 0.025	147.9 ± 17.3	0.519 ± 0.028	98.4 ± 3.0	**0.574** ± 0.100	**1.263** ± 0.222	**0.614** ± 0.050	**1.485** ± 0.190
C2	0.589 ± 0.029	126.2 ± 28.0	0.572 ± 0.029	99.6 ± 2.3	0.355 ± 0.066	0.586 ± 0.092	0.149 ± 0.017	0.359 ± 0.057
D	0.374 ± 0.089	134.8 ± 23.5	0.618 ± 0.034	99.5 ± 0.9	0.241 ± 0.073	0.381 ± 0.120	0.099 ± 0.036	0.258 ± 0.081
F	0.592 ± 0.036	175.6 ± 37.6	0.539 ± 0.035	97.9 ± 1.7	0.469 ± 0.140	0.845 ± 0.295	0.162 ± 0.030	0.379 ± 0.112
G	0.495 ± 0.047	155.4 ± 10.8	**0.464** ± 0.023	**93.6** ± 0.8	0.302 ± 0.074	0.624 ± 0.164	0.205 ± 0.027	0.657 ± 0.053
H	0.435 ± 0.050	167.7 ± 23.3	0.651 ± 0.019	96.4 ± 1.5	0.331 ± 0.092	0.604 ± 0.223	0.160 ± 0.038	0.423 ± 0.137
L	0.477 ± 0.046	196.1 ± 44.8	0.554 ± 0.023	95.0 ± 1.8	0.346 ± 0.209	0.679 ± 0.371	0.124 ± 0.032	0.410 ± 0.241

CV for ROI mean values of DTI metrics (CV_{mean}, see Equation 4.1. The distribution of CV_{mean} values for FA TractSeg ROIs was similar for most imaging sites, except for centre A, which had a higher average CV ($CV_{mean} = 5.4\% \pm 4.8\%$) than any other centre ($\text{CV}_{mean} \approx 3.2\% - 3.7\%$). Images collected on GE scanners showed a slightly lower average variation ($CV_{mean} = 4.7\% \pm 1.1\%$) than scans acquired on Siemens scanners ($CV_{mean} = 4.9\% \pm 3.7\%$). Centre A showed higher variation alone than when combined with other centres. Variation was highest when pooling all data together ($CV_{mean} = 5.8 \pm 2.3$; Figure 4.9 top left). A similar pattern was observed for TractSeg MD variations, where centre A was again showing a widely spread distribution of CV scores. Centres A ($CV_{mean} = 4.9\% \pm 3.7\%$) and F ($CV_{mean} = 3.7\% \pm 1.6\%$) had a higher variation than the other centres ($CV_{mean} \approx 2\% - 3\%$). Although average CV values for inter-scanner datasets seemed elevated, the effect was not as prominent as for FA metrics (Figure 4.9 top right). Overall variation was less for MD than FA metrics.

JHU CV. Generally, a higher intra-scanner variation ($CV_{mean} \approx 4.5\% - 7\%$) was observed for mean FA values within JHU ROIs than for TractSeg ROIs. Although merging datasets from different scanners increased the variation slightly with respect to some single centre data ($CV_{mean} \approx 5.5\% - 6\%$) the effect was not as prevalent as for TractSeg ROIs. Highest variation in single scanner dataset was again detected in centre A ($CV_{mean} = 6.7\% \pm 2.5\%$; Figure 4.9 bottom left). Deviation between means of MD within JHU ROIs was similar to that observed for TractSeg ROIs ($CV_{mean} \approx 2.5\% - 4\%$) Variation of MD on combined data from all GE scanners ($CV_{mean} = 3.4\% \pm 0.9\%$) was comparable to the variation on single GE scanners ($CV_{mean} \approx 2.5\% - 3.5\%$), and lower than for Philips ($CV_{mean} = 4.3\% \pm 1.8\%$) or Siemens ($CV_{mean} = 4.1\% \pm 1.1\%$) scanners (Figure 4.9 bottom right).

Template Space. After tensor-based spatial normalisation the voxel-wise CVs were computed for intra- and inter-scanner data. Subsequently, the average of the CV maps within JHU atlas ROIs were calculated (CV_{voxel}, see Equation 4.2). The mean and standard deviation for the 20 different mean CV values within JHU regions are presented in Table 4.7. Most obvious was the much higher average CV for FA imaged from Philips scanners (CV_{voxel} = 24.7% - 27.2%) compared to Siemens (CV_{voxel} = 16.1%-18.5%) and GE (CV_{voxel} = 14.7% - 17.2%). This discrepancy was also observed when comparing average CV scores of inter-scanner data for each of the three vendors. While data scanned on different Siemens or GE scanners showed a similar level of variation ($CV_{voxel} = 18.2 \pm 2.9$ and $CV_{voxel} = 16.3 \pm 2.3$, respectively), the average CV on Philips scanners was much higher ($CV_{voxel} = 27.9 \pm 8.1$). Pooling data together across vendors (All or Siemens & GE) showed that variation is settled

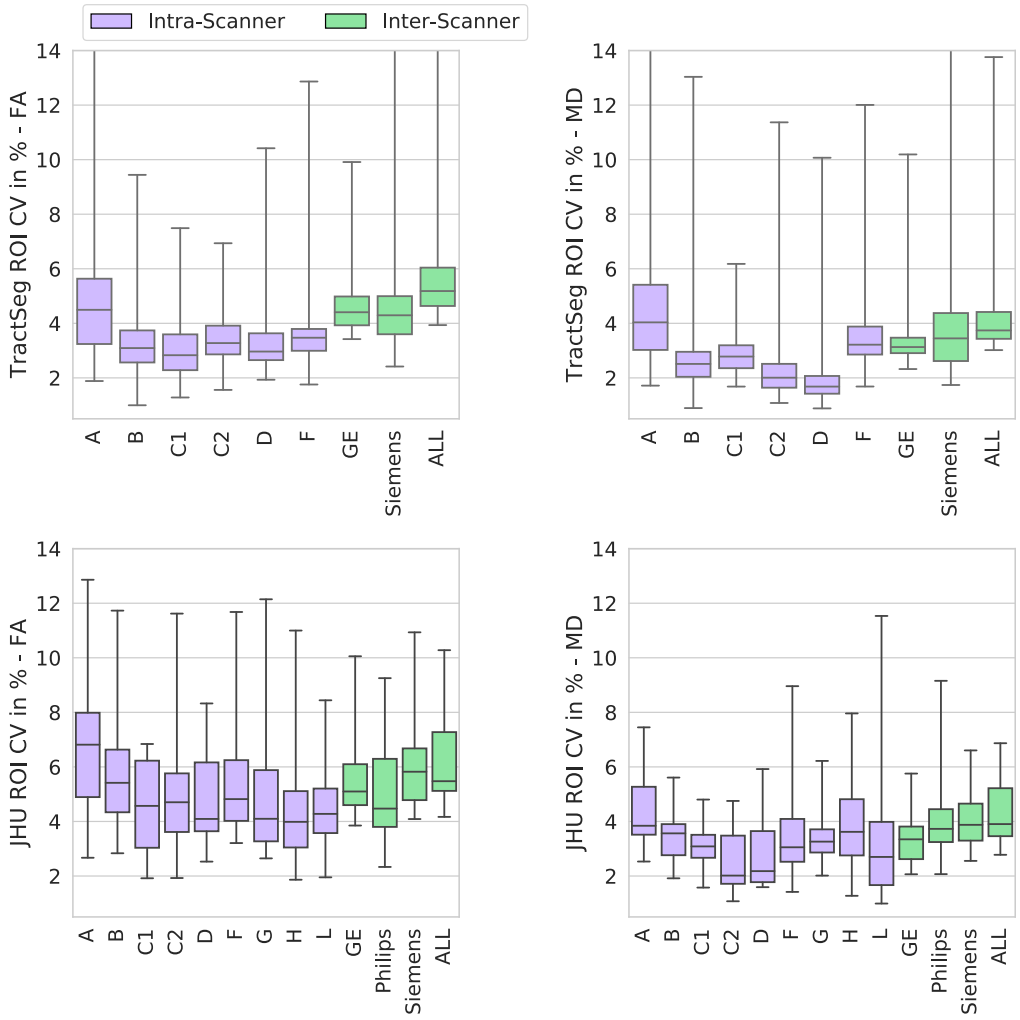


Figure 4.9: Intra- and Inter-Scanner Distribution of CV within ROIs for CENTER-TBI Imaging Sites. Each boxplot represents the variation of all ROIs of the particular atlas. **Top left:** Most centres/scanners showed a similar variation of FA mean values within TractSeg ROIs, averaging around $\approx 3\% - 4\%$, however, centre A showed an elevated average variation ($> 4\%$) compared to all other centres. Variation across sites was found to be substantial higher, and was comparable for intra-vendor (GE or Siemens) and inter-vendor (ALL: GE & Siemens combined). Data collected on GE scanners showed tendentially less variations than that acquired on Siemens scanners. **Top right:** MD variation was observed to be highest for the single centre A. Generally intra-scanner variation was on average lower than for pooled data from different scanners. Variation on GE scanner was lower on GE scanners than on Siemens scanners. **Bottom left:** Overall higher intra-scanner variation ($\approx 4\% - 7\%$) was observed for mean FA values within JHU ROIs than for TractSeg ROIs. Although merged datasets from different scanners increased the variation the effect was not as prominent as for TractSeg ROIs. **Bottom right:** MD variation was overall similarly distributed for intra- and inter-scanner variations. Whiskers show the full range of the distribution.

in between the variation measured for data collected on single-vendor scanners. Highest FA variation for the combined dataset from Siemens and GE scanners was found in the left hippocampal cingulum (ROI #6, CV_{voxel} = 22.3%). This was coherent with the highest variation in Siemens data for the same region (CV_{voxel} = 24.4%). Maximum FA for GE data were observed within the left ILF (ROI #12, CV_{voxel} = 20.0%). Lowest FA variability for GE and inter-vendor data (Siemens & GE) were found in left CST (ROI #2, CV_{voxel} = 12.4% and CV_{voxel} = 14.1%, respectively). Minimum variation for Siemens data were measured in the right CST (ROI #2, CV_{voxel} = 14.0%).

Variation for MD was observed to be lower (CV_{voxel} = 5.9% - 12.6%) than that for FA, which is in agreement with the previous analysis of regional variation in native space. Mean diffusivity variability was highest for Philips data, although the difference to other centres was not as drastic (centre A showed high CV for MD as well). For GE and inter-vendor data, the highest MD variability was observed in the forceps major (ROI #8, Siemens: CV_{voxel} = 15.3%, GE: CV_{voxel} = 18.6%, Siemens & GE: CV_{voxel} = 17.7%). The left temporal SLF showed the lowest for all three inter-scanner comparisons (ROI #18, Siemens: CV_{voxel} = 6.4%, GE: CV_{voxel} = 5.1%, Siemens & GE: CV_{voxel} = 6.0%).

Generally, data from A showed a higher variation for FA and MD than other Siemens scanner data. Inspecting data from Philips scanners visually showed more inadequately deformed fibre tracts (e.g. CC showed a more unusual shape). This seemed to indicate a less accurate spatial normalisation for Philips scans than for data from both other scanners, which might explained the elevated variation.

Figure 4.10 presents the CV maps of the DTI maps for the CENTER-TBI controls after spatial normalisation. For comparison it includes two maps for the centres with the highest (A) and the lowest (C2) variations (among Siemens and GE scanners). Philips scanners were not included in this analysis due to potentially insufficient registration to the study-specific template (based on visual examination). When including more subjects the maps become inherently smoother (averaging more images reduces single voxel outliers) and variation is conjointly reduced. Nonetheless, lower levels of variation were observed for single- than multi-scanner data. Overall variation was found to be higher in cortical GM areas (up to 80%) than in WM fibre tracts on FA maps (top panel: FA). Lower CV scores were detected for centre C2 than for centre A. This also translates to multi-scanner intra-vendor data (Siemens, GE) where GE showed lower CV scores than Siemens. This is particularly visible in cortical regions and to a certain extent in WM areas. While intra-vendor variation is likely an effect of different DWI acquisition and hardware biases of different scanners, the generally higher CV in cortical areas is partly also caused by more anatomical variation across subjects, that cannot fully be eliminated by spatial normalisation. Pooling data from

Table 4.7: Average of CV Maps within JHU ROIs. CV displayed as mean ± std in %

	Centre	Vendor	FA	MD
Intra-Scanner Intra-Vendor	A	Siemens	18.5 ± 3.3	10.2 ± 1.8
	B	Siemens	16.1 ± 2.3	7.4 ± 1.8
	C1	GE	15.6 ± 2.3	6.8 ± 2.4
	C2	GE	14.7 ± 2.3	5.9 ± 2.3
	D	Siemens	17.2 ± 3.5	7.8 ± 2.2
	F	GE	16.1 ± 2.5	8.0 ± 3.1
	G	Philips	24.7 ± 7.6	9.1 ± 3.7
	H	Philips	27.2 ± 7.8	12.6 ± 6.4
	L	Philips	25.8 ± 6.4	11.0 ± 7.4
Inter-Scanner Intra-Vendor	A, B, D	Siemens	18.2 ± 2.9	9.3 ± 2.0
	C1, C2, F	GE	16.3 ± 2.3	7.6 ± 2.8
	G, H, L	Philips	27.9 ± 8.1	12.8 ± 6.8
Inter-Scanner Inter-Vendor	A, B, C1, C2, D, F	Siemens & GE	17.6 ± 2.5	8.9 ± 2.4
	All	All	24.5 ± 6.3	12.2 ± 5.4

Siemens and GE scanners together yielded CV levels in between that of both single-vendor CV maps. A more detailed view of the FA displays lower variation (deeper blue) in the forceps major of the CC for C2 than for A. This tendency could also be observed for the CV of GE data compared to that of Siemens data (middle panel: FA-CC). These observations could also be quantified by counting the number of voxels within the brain mask that had higher CVs for one dataset than the other. This uncovered that 64% of the voxels had higher CVs on Siemens than on GE. Combining both datasets together (Siemens & GE) resulted in 71% or 42% of the voxels in a higher CV compared to GE or Siemens data, respectively. This further underlined the fact that a higher degree of variation is observed in Siemens data. Similar results were observed for CV maps of MD (bottom panel: MD). Data collected at centre C2 showed a much lower overall variation (deeper blue) than data acquired at centre A. Particularly noticeable in the WM, this lower variability was also visible when comparing inter-scanner data from single scanner manufacturers (Siemens vs. GE). Coefficients of variation were higher on Siemens than on GE scanners in 66% of the voxels within the brain. Pooling together data from both vendors resulted in increased variation compared to GE (73% voxels showed higher CV scores for Siemens & GE than for GE alone), however, in a lower variation with respect to Siemens data (43% voxels had a higher CV in Siemens & GE data than in Siemens alone).

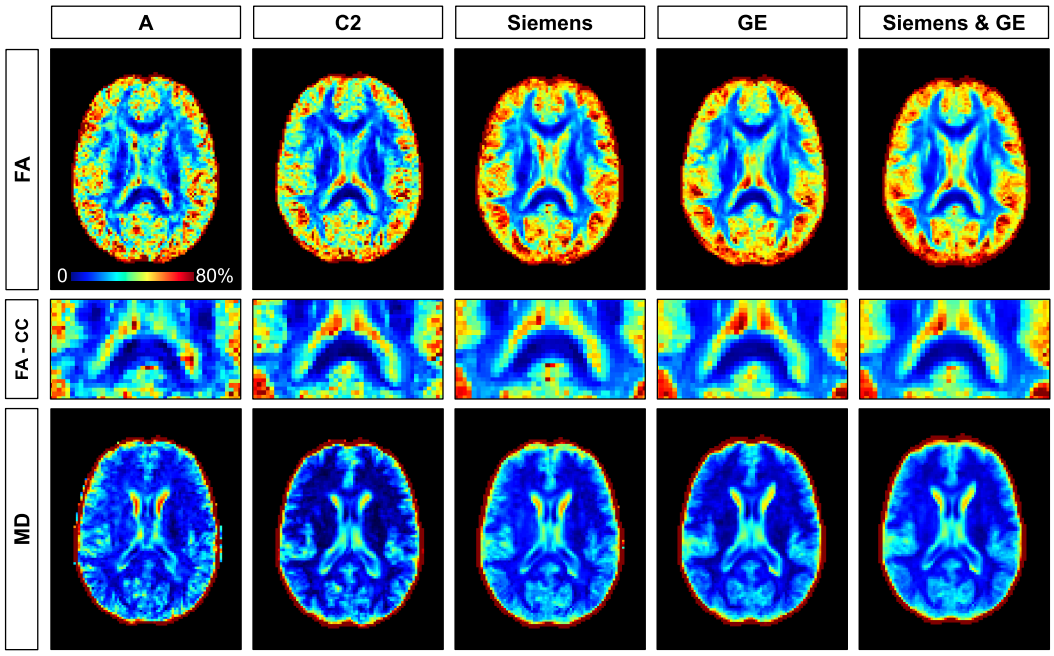


Figure 4.10: Coefficients of Variation Maps for FA and MD of CENTER-TBI Controls. **Top row:** CV maps for FA show higher variability for inter-scanner than in single scanner data. FA metrics derived from scans collected on GE had lower variation than Siemens data, which was visible for both single scanner (A vs. C2) and multi-scanner (Siemens vs. GE) comparisons. **Middle row:** Less variation was for example found on FA maps within the forceps major of the CC (FA - CC) for GE data. **Bottom row:** This trend was also present for MD maps, with lower CV scores for GE scans (C2 and GE). Colourbar is the same for all plots.

4.4 Discussion

4.4.1 Parcellation of Structural Brain Scans

Although partly robust, regional volume estimation via MALP-EM also deviated for data acquired on different scanners. Significant differences were found for both total brain and WM volume. This was even the case when starting with the exact same brain mask. Brain parcellation with MALP-EM is dependent on spatial alignment and deformable registration of multiple atlases and intensity based refinement. Both might be influenced by head orientation, SNR levels and voxel resolution, which makes the algorithm's robustness dependent on the image acquisition parameters. For the Scan-Rescan database, volume differences were subtle, and a 1-2% variation for volumes of tissue compartments across scanners seems to be a fair limitation for an automated process, as even variation between two manual annotators cannot be avoided completely [115, 210]. Nonetheless, this might be more pronounced for scans collected under different protocols or on scanners from different vendors. While travelling heads are not always feasible for large multi-centre studies, a quality check

of volumes derived from age and sex matched controls should be conducted prior to any analysis. Some of the regions found to be significantly different across scanners had deviating volume estimates for symmetric ROIs in both hemispheres. However, the affected ROIs were not of particularly large or small volumes, which may indicate a size independent systematic bias. Furthermore, the findings suggest that matching intensity spaces across scanners, could result in a positive effect diminishing volume discrepancies. However, this effect was only observed when intensities were matched across scanners with paired T1w scans. Since projecting all scans to the intensity space of an independent database (Cam-CAN) did not achieve the same result, the applicability for multi-centre studies will need to be investigated further. This holds especially true for patient cohorts with visible lesions on T1w images, as this might skew the intensity matching leading to a disadvantageous result. Simply computing relative region volumes with respect to the total brain volume showed that that discrepancy can partially be eliminated. However, this was only demonstrated for a fairly small set (12) of T1w scans of matched healthy controls. Normalising volume estimates to the whole brain size may be skewed if pathological patterns affect brain extraction. Regions of the MALP-EM atlas are directly adjacent to one and another. So, if one region is segmented more, it almost always means that another ROI segmentation needs to shrink. Future experiments could aim to understand the interplay between regions that were found to have significantly different volumes.

4.4.2 Segmentation of White Matter Tracts

Despite using a harmonised diffusion protocol, significant volume differences were observed when comparing TractSeg segmentations on different scanners, using the Scan-Rescan database. One reason for that could be the model based nature of the algorithm. TractSeg is a neural network that was trained in a supervised way to predict a voxel's association to a specific tract based on the peak of the fibre *orientation distribution function* (ODF). Although TractSeg has been shown to outperform many other segmentation approaches [263], its superiority was shown on test data from the same cohort acquired for the HCP. Since machine learning models are susceptible to domain shifts as they are present for different MRI datasets, predictive segmentation might not generalise well to unseen data. Since TractSeg's neural network was trained on HCP, it also learned to identify/segment fibre tracts based on intensities as they were present in the HCP dataset. For diffusion images that were acquired differently or on a different scanner, intensities for the same fibre tract may deviate from that in the HCP database. This would result in a TractSeg failing to segment WM tracts adequately. This became obvious when examining the TractSeg region volumes on differently acquired diffusion weighted scans (Multi-Acquisition dataset).

Protocols with higher angular resolution improve the accuracy of the computed ODF [279], which directly impacts the peaks of the ODF and with-it the segmentation performance of TractSeg. Although the Multi-Acquisition dataset consisted of 60 diffusion weighted images for all five different protocols, the acquisitions were distributed on a different number of shells. It is not entirely clear how much impact the different number of shells had on computing the ODF's peaks, however, TractSeg seemed to be biased towards the data it was trained on. This observation was underlined, when aiming to parcellate data for CENTER-TBI. While decent results were obtained for data collected on Siemens and GE scanners, parcellation was inadequate for data acquired on Philips scanners. This was consistent for Philips scanners at different imaging sites, suggesting a vendor-specific negative bias. Training data for TractSeg was exclusively collected on a Siemens scanner. Why TractSeg performed better on data from GE than Philips scanners may be part of future investigations. Volumes computed for tracts from the JHU atlas seemed to be affected less by different protocols. This may be explained by the underlying method, based on registration, to segment WM fibre tracts. While registration can be affected by different intensities, it is less sensitive to variability in the data - when choosing an appropriate cost function - than predictive models. This holds especially true for registration of FA maps, as these are normalised metrics, normally ranging between [0, 1]. By definition there is a high contrast between WM tracts and GM tissue on FA maps. A segmented fibre tract that also unintentionally includes GM voxels, possibly exhibits a higher variation than more precisely segmented tracts only including WM. The fact that FA variability was overall higher for JHU ROIs ($> 4\%$) than for TractSeg ROIs ($\approx 3\%$), could be a consequence of this segmentation inaccuracy. Comparing JHU and TractSeg against each other is difficult, as there set of ROIs is different and same regions may not overlap entirely. Nonetheless, this observations are coherent with TractSeg being reported to be superior to single atlas segmentation [263].

4.4.3 Variability of DTI Metrics

Variation of MD was consistently found to be lower than that for FA metrics. This robust finding was observed for both the TractSeg and JHU analysis and could be replicated for all different datasets (Scan-Rescan, Multi-Acquisition and CENTER-TBI database). A lower MD variability agrees with previous studies and makes this result highly generalisable. It has been hypothesised that individual variability in eigenvalues of the diffusion tensor may compound when combined to calculate FA metrics [72]. The lower contrast seen on MD maps leads to an inherent lower variation within the scan itself (and also lower impact of miss-registrations and -segmentations). Particularly WM regions usually appear to have mostly homogeneous MD. This became also obvious when inspecting the voxel-wise

CV maps. These displayed mostly equal CV values throughout the brain for MD metrics, whereas in contrast FA maps showed a strong divergence between GM and WM regions. This is in line with the previously reported higher FA variation in GM than in WM [256], and is likely a consequence of better congruence of anatomical WM structures. While central fibre tracts, such as the CC, can be more easily aligned or segmented, peripheral endings of fibres within GM vary much more in shape and orientation. Interestingly, when considering the number of ROIs with significant differences, more were counted for MD than for FA maps. A higher number of ROIs with inter-scanner divergence for MD than for FA has been found beforehand [66].

Furthermore, the inter-scanner variation was similar to the highest observed intra-scanner variation. For example, FA CV scores for combined GE data (CV = 16.4%) was of comparable to the CV of GE data from centre F (CV = 16.1%), but increased with respect to centre C1 ad C2 (CV = 15.6% and CV = 14.7%, respectively). For inter-vendor analysis the total variation (Siemens & GE CV = 17.6%) was in between the intra-vendor variation (Siemens CV = 18.2%, GE CV = 16.3%). Coefficients of variation are based on average values across datasets, thus, combining datasets with different levels of variation mostly resulted in CV scores that were settled in between (compare values Table 4.7). Additionally, including more data can also smooth out natural differences in anatomy among subjects leading to reduced noise and eventually lower variability. This observation is further supported by the previously reported decrease in variability through image denoising via median filters [164]. There was an evident bias between scanners in the CENTER-TBI database, where lower variability was found for GE than for Siemens scanners. This is cohesive with aforementioned findings [283].

Despite the total number of unique gradient directions being the same across the MRI protocols of the Multi-Acquisition database, they were distributed on one to five shells of different b-values. This had an effect on both FA and MD metrics. The most distinct difference was found for the single-shell data with 60 gradient directions in comparison to the multi-shell data. One reason could be that multi-shell imaging schemes show a higher reproducibility than single-shell data [245]. This result, however, was produced for multi-shell data with $b \geq 1000\, s/mm^2$, whereas the data presented here relied on *inner* shell b-values ($b \leq 1000\, s/mm^2$). Future experiments may need to consider number of shells and directions as two separate variable factors. Possibly, splitting different acquisitions into their different shells may help to build a more descriptive model. Furthermore, it has been shown that the choice of parameters may impact FA and MD maps unequally [206]. Understanding each parameter's contribution to DTI variability will help to design more robust MRI protocols for multi-centre studies and foster solutions for diffusion MRI harmonisation.

Regarding multi-centre data, the findings suggest that even with quasi harmonised protocols, MR image acquisition will be affected by hardware settings (scanner model and vendor etc.) and scanner operators (different magnitude of head motions at different centres). Nonetheless, one should aim to collect data under a harmonised protocol to minimise chances for variation. Specifically, for CENTER-TBI data, Philips scanner data will need to be examined closely to understand its discrepancy to Siemens and GE data.

4.4.4 The Difficulty of Measuring Variability

The experiments presented here have shown variability for brain region segmentation and DTI metrics. Although this is expected to a certain extent, quantifying differences between data can be challenging. Any region based analysis will only be as good as the region segmentation. White matter parcellation via TractSeg, for example, was unsuccessful for data collected on Philips scanners for CENTER-TBI. While the discrepancy for the given data was obvious and the systematic error was easy to detect, this might not be the case for more subtle differences. A bias of volume segmentation for measuring regional variability of DTI maps might be inevitable, but should be closely monitored throughout the analysis. For the Multi-Acquisition database no strong correlation between volume variation and FA or MD variation was found. However, some outlier regions were detected that showed high variation for both volumetric and DTI measurements. These regions were the ones with smallest segmented volumes. The effect of high variation in smaller regions has been reported before [193, 256]. This seems characteristic for low volume ROIs as small deviations, due to scan differences or flawed segmentation, strongly impact CV scores. One way to avoid mistakes from methodological biases is to examine the problem from different angles. Besides the ROI analysis, a voxel-wise approach was chosen to estimate reproducibility of diffusion data. This facilitates the identification of robust results, such as the previously mentioned lower variability for MD than for FA, but also the detection of problematic data (i.e. here Philips MR scans). Most previous studies for reproducibility rely on CV scores to estimate the differences between data. However, comparing CV scores of mean values within ROIs (CV_{mean}) against the CV values on a voxel-wise level (CV_{voxel}) showed very different magnitudes. This has been reported previously [241], strongly suggesting the dependence of CV values on how they were measured. Therefore, it might be more important to examine the change of variation when pooling data together, rather than accepting a cut-off threshold for acceptable reproducibility (e.g. 10% as suggested by Morenco et al. [168]). Overall, the experiments focused on assessing variation in different datasets. Future investigation will need examine more closely, how these differences are caused. The findings showed that IDPs vary for different MRI acquisition parameters or imaging hardware. These

factors and subject-specific confounding effects (e.g. head motion) jointly influence the image signal. Disentangling the contributions of individual factors, however, is complex. Statistical regression models will need to incorporate QC variables and other confounding factors into any analysis.

4.5 Chapter Summary

In this chapter different datasets were analysed to estimate the reproducibility of brain region segmentation as well as the variability of DTI metrics. Discrepancies between ROI segmentation were found for both the parcellation of T1w and DTI scans. The former, based on MALP-EM, shows regional differences which can be partially reduced by using a robust mask[9] or by matching intensity domains. However, further investigation is needed to assess the applicability of such methods to more complex multi-centre data. White matter parcellation through TractSeg seemed less replicable for data collected on different scanners or different MRI acquisition protocols. Tract segmentation via JHU atlas appeared to be more robust, but also less precise. Differences for DTI metrics were region-specific and highlighted the non-linear inter-scanner variability. A direct connection between parameters such as number of gradient directions per shell and number of shells could be drawn, but will need further experiments to gain a deeper insight. Variation of DTI metrics was also dependent on the employed scanner's vendor. Overall, a lower variability for MD than for FA was measured consistently in all experiments on various datasets. Inter-scanner variability may impede analysis for combined multi-centre databases, in particular if variation is high for individual scanners. Complexity of those databases will need to be addressed with advanced harmonisation techniques and statistical modelling. Although these techniques can mitigate effect, they likely will not remove all biases. Generally, more data may help to regress out confounding factors and the more similar the cohorts and acquisition parameters across sites the better. Without doubt, databases will need to be carefully examined to understand hidden stratification [189] to allow drawing sensible conclusions from any experiment.

[9] for experimental purposed the same mask for different scans was used

Chapter 5

Harmonisation of DWI for Multi-Centre Studies

5.1 Introduction

5.1.1 Sources of Variation in Diffusion MRI

Acknowledging the benefits of larger databases, multi-centre studies have been gaining increasingly more attention over the years. However, as seen in previous chapters, diffusion weighted MRI are dependent on scanning parameters as well as imaging hardware. Diffusion MRI signal is influenced by the number, direction, strength and duration of different gradient pulses. Regional FA differences have been reported when comparing different b-values [24]. In particular larger b-values have been associated with reduced DTI measurements [194]. Besides their magnitude, the number of b-values used has also been shown to affect the diffusion signal. For example, a multi-shell acquisition was found to be beneficial to reduce Rician noise related biases [41]. Moreover, a higher number of diffusion directions was shown to improve contrast between GM and WM on FA maps, but had little effect on MD metrics [80]. A more elaborate study showed empirically that a higher angular resolution improved SNR and recovery of the orientation density function [279]. Interestingly, different diffusion parameter maps required more gradient directions to achieve a near-optimal SNR than others (e.g. FA: 62-66 vs. MD: 58 directions) [279]. Furthermore, spatial resolution (i.e. voxel size) has been reported to affect DTI metrics. A larger voxel size was linked to increased MD values, but decreased FA measurements [194]. Moreover, scanner hardware such as, for example, field strength of the magnet [101] also have an impact on image quality. While some sources of variations can be avoided by a careful study design (e.g. using same

acquisition protocol), some others are inevitable. To address this problem, harmonisation of diffusion MRI data has become a more active research field in recent years.

5.1.2 Spherical Harmonics & Rotation Invariant Features

Some harmonisation techniques operate in the DWI space, while some others employ SHs[1] [129] as an alternative representation. Generally, these are a mathematical notation to represent any spherical function $f(\theta, \phi)$ as a sum of its harmonics:

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} C_{l,m} Y_{l,m}(\theta, \phi) \tag{5.1}$$

with $Y_{l,m}(\theta, \phi)$ representing the SH basis function of order l and degree m, and $C_{l,m}$ are the corresponding SH coefficients. The energies of the SH coefficients for each order l form a set of *rotation invariant spherical harmonics* (RISH) features:

$$\|C_l(f)\|^2 = \sum_{m=-l}^{m=l} (C_{l,m})^2 \tag{5.2}$$

Orders of SH are even numbers only. Hereafter, only the highest order will be mentioned to indicate the computation of SH orders up to that maximum order (e.g. l_{max}=4 means order 0, 2 and 4 were computed).

5.1.3 Related Work

Meta-Analysis. A classic approach to boost sample size is the meta-analysis that spans across imaging centres and studies. For this, group-wise analysis is at first performed independently for each site and findings are then statistically examined to find consistent results across sites. Alternatively, data are analysed together by reducing variation through metric standardisation and modelling site-specific biases as random effects [198] (also see Chapter 3). Ideally, all image processing steps should be the same to minimise variability introduced through unequal data handling [222]. A simple approach is Z-scoring, that standardise images by transforming the signal to a standard distribution (mean $\mu = 0$ and standard deviation $\sigma = 1$) for each site individually [70]. However, this is highly dependent on the subject cohort at each site and may be too crude to account for non-linear changes across different brain regions. More recently, new approaches have been introduced, which will be summarised in the following paragraphs. The interested reader is further referred to the review published by Pinto et al. [198].

[1] reminder: spherical harmonics

ComBat. The *combined association test* (ComBat) was initially introduced for gene expression analysis [117], but was later adapted for diffusion MRI harmonisation by Fortin et al. [70]. Hereby, differences across scans and sites are estimated by modelling the diffusion signal as an adjustment of intensity location (additive factor) and scale (multiplicative factor). The underlying assumption is that both factors follow a parametric distribution that can be retrieved by an empirical Bayesian model to reduce site-specific effects. A diffusion metric y can be represented for each voxel v in scan n from site i as:

$$y_{i,n,v} = \alpha_v + \mathbf{X}_{i,n}\,\boldsymbol{\beta}_v + \gamma_{i,v} + \delta_{i,v}\,\varepsilon_{i,n,v} \tag{5.3}$$

where α_v is the overall signal for voxel v with residual error $\varepsilon_{i,n,v}$. Covariates of interest (e.g. sex, age) are represented as $\mathbf{X}$ with a corresponding vector of regression coefficients $\boldsymbol{\beta}_v$. The terms $\gamma_{i,v}$ and $\delta_{i,v}$ stand for the additive and multiplicative effects, defined as prior Gaussian and Inverse-gamma distributions, respectively.
Once the distributions of hyper-parameters are estimated ($\hat{\alpha}_v$, $\hat{\boldsymbol{\beta}}_v$, $\hat{\gamma}_{i,v}$, $\hat{\delta}_{i,v}$), the harmonised signal $y^*_{i,n,v}$ can be calculated as:

$$y^*_{i,n,v} = \frac{y_{i,n,v} - \hat{\alpha}_v - \mathbf{X}_{i,n}\,\hat{\boldsymbol{\beta}}_v - \hat{\gamma}_{i,v}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + \mathbf{X}_{i,n}\,\hat{\boldsymbol{\beta}}_v \tag{5.4}$$

The advantage of this method is the possible application to any diffusion parameter map (e.g. FA, MD) and in fact has also been successfully applied to harmonise non-diffusion metrics such as cortical thickness [69]. Due to its voxel-wise modelling of site-specific effects, ComBat can account for local scanner differences. It has been shown to outperform other strategies based on voxel-wise linear regression, such as RAVEL (removal of artificial voxel effect by linear regression [71]) or surrogate variable analysis [70].
However, ComBat assumes that location and scaling factors are represented by specific parametric prior distributions, which may not generalise for all diffusion measurements [126]. Furthermore, ComBat is applied to DTI parameter maps rather than the DWI signal. The performance of pre-processing steps could be biased by the differences of the unharmonised images. According to Karayumak et al. [126], this could lead to site-specific effects being latently propagated through the processing pipeline, and make it harder to detect and correct for those effects. Furthermore, it has been shown, that ComBat possibly has an adversary effect on between-group differences if data were processed slightly differently [33]. This, however, can be mitigated by applying the same processing pipelines.

Linear Scaling of Spherical Harmonics. A different approach was suggested by Mirzaalian et al. [177], for which diffusion MRI data were first parcellated[2] and converted to

[2]FreeSurfer parcellation of T1w images, which were projected to diffusion images.

their SH representation (Section 5.1.2). Next, a regional linear mapping between the RISH features was calculated to map SH images from one acquisition site to the image domain of the reference site. After scaling the images region-wise, SH coefficients were adjusted for each voxel within the brain. Later on, this approach was improved by sub-dividing the regions into smaller segments and refining this finer parcellation on basis of RISH feature intensities [178]. To circumvent the sub-optimal parcellation (i.e. registration errors for too large ROIs) entirely, the same concept of SH scaling was embedded into a registration-based framework [179]. For this, a study specific template was calculated via multi-modal registration of RISH feature maps. For each site and each RISH feature map the expected value was calculated from the spatially normalised images:[3]

$$\mathbb{E}^{\,l}_{\,i}(v') = \frac{1}{N_i}\sum_{n=1}^{N_i} \left\|C_l(v')\right\|^2_{i,n} \tag{5.5}$$

where $\left\|C_l(v')\right\|^2_{i,n}$ is the value of voxel v' in the RISH feature map of order l for the n^{th} subject at site i within template space. Here, v' corresponds to the voxel v in subject native space, such that $v' = \Psi_n(v)$, with Ψ_n defining the subject's diffeomorphic deformation field that projects image from native to template space. The scaling map was then defined per RISH feature map:

$$\mathfrak{S}_l(v') = \sqrt{\frac{\left\|C_l(v')\right\|^2 + \mathbb{E}^{\,l}_{\,ref}(v') - \mathbb{E}^{\,l}_{\,src}(v')}{\left\|C_l(v')\right\|^2}} \tag{5.6}$$

For new DTI scans, RISH feature maps are computed and registered to the template. This allows the backprojection of the scaling maps in order to map SH components from the source site to the reference site domain in native space:

$$\hat{f}(v,\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{m=l} \mathfrak{S}^{\psi}_{l}(v)\, C_{l,m}(v) Y_{l,m}(\theta,\phi) \tag{5.7}$$

with $\mathfrak{S}^{\psi}_{l}$ representing the backprojected scaling map $\Psi_n^{-1}(\mathfrak{S}_l)$.
Karayumak et al. [126] redefined the scaling map for an independent study as:

$$\mathfrak{S}_l(v') = \sqrt{\frac{\mathbb{E}^{\,l}_{\,ref}(v')}{\mathbb{E}^{\,l}_{\,src}(v') + \varepsilon}} \tag{5.8}$$

with ε representing a very small number to avoid division by zero. Being model-independent, the calculated site-specific mapping can easily be applied to any other data from the same cohort site. Hereby, the scaling of SH coefficients allows the harmonisation of the signal amplitude without altering the principal diffusion directions [177]. In addition, linear scaling of SH coefficients has been reported to harmonise the signal more accurately within most

[3] spherical function term (f) is dropped for brevity

of the WM fibre tracts than ComBat [126]. Once the diffusion signal had been harmonised, any desired downstream analysis can be applied. However, this method also needs matched controls across sites (ideally the same subjects scanned repeatedly) scanned with relatively similar acquisition parameters. The voxel-wise scaling is dependent on spatial normalisation, which can be time-consuming in practice, especially for the required multi-model RISH feature map registration.

Non-Linear RISH Feature Regression. Besides computing linear scaling maps, machine learning methods were used to learn non-linear functions in order to harmonise DTI or RISH feature maps across databases [127]. The employed models were a RF regressor [26] to scale either DTI (RF-DTI) or RISH feature maps (RF-RISH) and a CNN [150] trained on RISH features (CNN-RISH). All these approaches required a direct voxel-wise correspondence of images from the same subjects acquired on two scanners. Comparing diffusion parameter maps from harmonised and reference data showed that all three non-linear regression models (RF-DTI, RF-RISH and CNN-RISH) outperformed the linear RISH feature scaling. Moreover, training a RF on RISH features seemed more beneficial than training on DTI maps. The CNN regression showed the best performance for all evaluation metrics [127].

Spherical Harmonics Mapping with Neural Networks. With the success of deep learning, many neural network approaches have been suggested to harmonise diffusion MR scans. In contrast to the previous RISH feature regression, most methods operate directly on the SH coefficient maps. Tax et al. [238] provided a benchmark dataset to challenge researchers to submit approaches for diffusion imaging harmonisation.[4] The harmonisation task required the mapping of DWI images from one scanner (Prisma) to another (Connectom), for which data of ten subjects imaged on both scanners was provided. The summary of the submitted results revealed a clear trend towards deep learning. While the neural network architecture and training strategies varied for different approaches, they had in common the use of pairwise 3D patches extracted from SH images matched on a subject- an voxel-wise level across scanners. The *spherical harmonic network* (SHNet) consisted of three fully connected layers operating on SH coefficient maps to learn the mapping between both scanners. Built up on that, the *spherical harmonic residual network* (SHResNet) included different convolutional pathways for each of the SH orders (i.e. $l = 0, 2, 4$). While both networks were trained on small patches (3×3×3 voxel) matched between scanners, the SHResNet also included *residual blocks* (ResBlocks) that combine the in- and output of the

[4] this challenge was hosted as part of MICCAI 2018, the 21st International Conference on Medical Image Computing and Computer Assisted Intervention

different pathways. Such residual connections are a common approach for neural networks to improve its performance [96] and in theory can be included multiple times. The authors found an optimal performance with two ResBlocks (with only marginal improved compared to the use of one unit) [135]. Similarly, Tanno et al. suggested a *fully-convolutional shuffling network* (FCSNet, also see [237]) that included four convolutional layers and one skip connection that concatenates the input to the first layer and the output of the last convolutional layer. In contrast to SHResNet, FCSNet operates on a much larger receptive field size (isotropic patches of 11 voxels) and consisted of only one single pathway for all SH orders ($l_{max} = 6$ or 8). Furthermore, the channel-wise loss function (same as for SHResNet, i.e. *mean square error* [MSE] between the raw input and harmonised output SH signal) was enhanced with a RISH feature based loss. Besides applying convolutions in SH image space, Koppers et al. also introduced a *spherical network*. This samples at first 30 equidistant gradient directions from the input SH signals and subsequently performs spherical convolutions before converting the signal back to original SH space [136]. Among the mentioned architectures, FCSNet outperformed all other neural networks on the *Multi-shell Diffusion MRI Harmonisation and Enhancement Challenge* (MUSHAC) benchmark database [238].

Scaling of DWI Data. Many previously mentioned methods operate in the SH or RISH feature space, however, direct harmonisation of DWI signal has been suggested as well. One approach is the *method of moments* (MoM) [104], which effectively is a voxel-wise linear scaling of DWI images from source (S_{src}) to reference space ($\hat{S}_{src}$), where the multiplicative factor α and the additive factor β are derived from the first spherical moment (M_1) and second central (C_2) spherical moment of the DWI data:

$$\hat{S}_{src} = \alpha S_{src} + \beta \quad with \quad \alpha = \sqrt{\frac{C_{2,ref}}{C_{2,src}}} \quad and \quad \beta = M_{1,ref} - \alpha M_{1,src} \tag{5.9}$$

The first and second moments correspond to the spherical mean and variance of the DWI signal, respectively. In practice, spherical moments are first computed for all subjects and then projected to a common space (spatial normalisation). For each individual site the median of the first and second moments are computed before the scaling factor maps are derived. These are then backprojected to the native space of each subject from the source site to eventually project the DWI signal to the reference site domain. The authors reported that MoM harmonisation preserved both the shape and directional information of the signal profile.

Representation Learning. A different research direction for data harmonisation aims to find a new representation of diffusion MR images that removes undesirable variation

across scanners, but preserving the underlying biological variability of interest. One suggested approach is the learning of a sparse dictionary [165], that consists of base elements D that can linearly be combined with weighting coefficients α_n to reconstruct any sample x_n in the dataset. The dictionary acts as sparse representation only keeping features necessary to describe a database.

$$x_n \approx D\alpha_n \tag{5.10}$$

The dictionary can be iteratively learned to adapt to training samples $X = \{x_1, ..., x_n\}$:

$$\arg\min_{D,\alpha_n} \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{2} \|x_n - D\alpha_n\|_2^2 + \lambda_i \|\alpha_n\|_1 \right) \tag{5.11}$$

whereas the ℓ_2-norm and ℓ_1-norm promote data similarity and sparsity of the coefficients $\boldsymbol{\alpha}$, respectively. St-Jean et al. [234] applied this concept to diffusion MRI harmonisation to implicit mapping between domains of a *source* and *reference* scanner. One option is to construct reference dictionary to encode the information in the data from the reference scanner. This can then be applied to data from the source scanner. The idea is that the reference dictionary reconstructs images that preserve only common scanner feature, while filtering out source-specific properties. This implicitly maps source data to the reference scanner domain. Alternatively, a dictionary can be jointly learned on data from multiple scanners such that the base elements represent only the common features (i.e. diffusion signal) and discard site-specific effects (i.e. scanner variations). Such a dictionary would project images from different sites to one common domain, rather than projecting one dataset to the space of another pre-selected reference scanner. For diffusion data harmonisation, St-Jean et al. [234] extracted 3D patches of a DWI volume and spatially corresponding patches from the angular neighbouring volumes to learn dictionary.

Similar in its core idea to learn a scanner-agnostic representation, Moyer et al. [182] has very recently developed a *variational auto-encoder* (VAE) framework to harmonise diffusion MRI data. This neural network consists of two parts: The *encoder* that compresses the input information to an intermediate representation, and the *decoder* that tries to reconstruct the input image data from the latent space variables. For data harmonisation the goal is to learn an encoding for the input images such that the latent representation is independent of the scanning site. The employed architecture consisted of two fully connected layers for each the encoder and decoder. These were trained on voxels and their immediate six neighbours sampled from from SH images ($l_{max} = 8$) for multi-shell data and one non-diffusion weighted b_0 volume. The network parameters were learned via the reconstruction loss for the SH signal (all seven voxels) and two auxiliary losses. These were the reconstruction error for the DWI signal (centre voxel only) as well as an adversary loss, which captures the potential

to distinguish samples from different sites. This approach has been shown to outperform a re-implementation of linear RISH feature scaling as previously suggested [179].

5.1.4 Advantages and Limitations of Existing Methods

While global scaling (Z-Scoring) is straight-forward to apply, there is little potential to remove region-specific biases across the brain scans. Harmonising data with ComBat provides the important advantage of simultaneously considering also clinical variables. A covariate matrix holding such information can individually be designed for any research questions. This also means that images from multiple sites can easily be harmonised all together. On the other hand, ComBat operates on diffusion parameter maps (i.e. FA and MD), meaning DWI scans were fully pre-processed before applying the harmonisation. Site-specific differences in acquisition could already lead to biases in data processing. For example registering data to a common template is dependent on diffusion image intensities. Data from a particular site may register less successfully to a template than others, due to different image properties (e.g. noise level).

With the initial success of RISH feature harmonisation by Mirzaalian et al. [177], many following approaches also used either SH or RISH features representation. One of the most promising techniques is the SH scaling via RISH feature matching in a registration framework [126, 179]. This has been shown to outperform ComBat [126] as well as other competitive methods (including neural networks) [238]. Since this harmonisation method operates on a voxel-wise level, it can account for regional differences. However, similar to ComBat[5] this relies on accurate spatial normalisation which is both time consuming and error-prone. Especially, as the suggested approach requires multi-modal registration. With increasing number of subjects, the computational burden increases, as each scan has to be individually spatially normalised. Moreover, in case of more than two imaging sites, one would need to select one reference site [32] and it is not entirely clear how to chose that. Although successful in both a ROI and registration framework, simple scaling of RISH models may not be able to capture non-linear intra-site differences. For example, if site-specific biases had a stronger impact on a subgroup of subjects than for others within a site, the data used to compute scaling maps might not adequately represented this subcohort. A single scaling map per site may not be enough to account for that, which is why non-linear regression models for RISH feature matching might have been superior [127]. The limitation of one scaling map per imaging site was later counted for by introducing an adaptive approach learning different scaling maps per subject. Hereby scale maps were computed from three subjects form the training data that were most similar (MSE between RISH features) to

[5] ComBat can also applied on a ROI level

the test subject. This could further boost the harmonisation performance, showing a strong improvement in comparison to competing algorithms [187]. Despite this, the dependency of spatial normalisation and selection of a reference site remains.

Tax et al. presented the performance of different algorithms on a benchmark dataset [238]. This revealed the superiority of linear scaling of SH (see above [126]) over most deep learning approaches. One explanation could be that the data may not be sufficient to train a successful mapping via neural networks, however, FSCNet showed an equivalent strong performance. One distinct difference between FSCNet and other neural network architectures presented (e.g SHResNet), is the receptive field size. While most approaches operated on small patches (3×3×3 voxels), FSCNet used large 3D patches (11×11×11 voxels), hence, including much more image semantic context. The benefits of a larger receptive field have been shown multiple times in other computer vision tasks [37, 64, 91]. Therefore, different architectures might be important, but cannot outweigh including enough information. However, the larger the patches the higher the computational burden and neural networks can be difficult to train adequately particularly with limited number of data. In addition, the neural networks presented also did not support direct multi-site data harmonisation.

Moyer et al. [182] addressed this shortcomings by deriving a site-unspecific representation via a VAE. This does not require any registration to template space and further allows model learning for multiple sites simultaneously. Imaging data from various sites can either be projected to a common *scanner-agnostic space* or to one selected site. The authors could show that this approach outperformed linear SH scaling as suggested by [179], however, it is not clear whether it would perform better than the adaptive scaling approach (see above).

5.1.5 Aims

Firstly, the linear RISH feature scaling framework as well as a neural network baseline (i.e. FSCNet) were re-implemented to test their harmonisation performance on a benchmark dataset. The aim was to understand whether multiple, individual network pathways for SH images of different orders are beneficial for diffusion data harmonisation. This was inspired by the SHResNet architecture described above, however, was trained under similar conditions as the FSCNet, which did not happen in the benchmark publication [238]. Secondly, two options to improve the scaling maps for linear RISH feature scaling for the CENTER-TBI database were explored. This included the denoising of the scaling maps as well as the subselection or weighting of images from available subjects to compute the scaling maps. Concepts were then applied to harmonise CENTER-TBI data from two different scanners and evaluated based on RISH features and DTI metrics. Finally, the impact of data harmonisation on mTBI subject analysis was examined.

5.2 Data & Methods

5.2.1 Databases

MUSHAC. The MUSHAC data provided by Tax et al. is a benchmark dataset for diffusion MR image harmonisation. Diffusion and structural images were acquired for ten healthy volunteers on a Siemens Prisma scanner (max. gradient 80 mT/m) and a Siemens Connectom scanner (max. gradient 300 mT/m). The DTI protocol was the same on both scanners and entailed 30 directions for two shells with $b = 1000$ and $3000\,s/mm^2$ and seven non-diffusion weighted images (b_0). The scans were collected through parallel imaging (GRAPPA=2) with $TE = 89\,ms$ and $TR = 7200\,ms$. The acquired voxel size was isotopic of $2.4\,mm^3$. All data were already corrected for susceptibility and eddy-current distortion, head motion and bias fields. Further, images from different scanners were aligned to each other for each subject individually. A brain mask, was computed on T1w images that was than inversely projected to DWI space via rigid coregistration of b_0 to T1w images. For more detail the reader is referred to the original publication [238].

CENTER-TBI. This database has been described previously (Section 2.2) and more information can be found online.[6] For this analysis the subset acquired at Cambridge was chosen, as it was the only site that scanned the same control subjects on two different scanners. All other centres either employed only one scanner or collected diffusion images of different healthy volunteers on each of their scanners. To analyse mTBI patients, data from the two week period post-injury were selected, as this provided the most patient scans. An overview of the data is given in Table 5.1.

5.2.2 Benchmarking Implementations of Harmonisation.

Among the previously mentioned algorithms, the linear RISH feature scaling as well as a deep learning approach similar to the *FSCNet* (Section 5.1.3) were re-implemented as these seemed most promising at the time of experiment setup. To benchmark the implementations the algorithms were applied to the openly available diffusion images from the MUSHAC database. For this analysis, the focus laid on the inner shell data ($b = 1000\,s/mm^2$), since the ultimate goal was to apply these concepts to the CENTER-TBI database, which collected single-shell diffusion MR images only. The DWI scans of the ten available subjects were transformed to SH representation (MRtrix3 `amp2sh`). Thereby, the default settings were used which did not normalise the data by the accompanied b_0 image, as initial tests

[6] `www.center-tbi.eu/project/mri-study-protocols`

Table 5.1: Overview of Cambridge Data for CENTER-TBI

		Prisma	Trio	Both
Controls	# Subjects/Scans	7/7	7/7	7/14
	Age at Scan	45 [32, 62]	45 [32, 62]	45 [32, 62]
	Sex (M/F)	5/2	5/2 F	10/4
Patients	# Subjects	14	18	32
	Age at Scan	51 [21, 62]	44 [21, 67]	47 [21, 67]
	Sex (M/F)	10/4	12/6	22/10
	DPI	21 [8,33]	17[11,26]	19 [8, 26]
	GCS (13/14/15)	1/1/12	0/3/15	1/4/27
	GOSE ≤ 8	8	8	16
	GOSE $= 8$	6	10	16

had shown strong outlier signal introduced by normalisation. The data allowed to estimate coefficients of SH up to order six (l_{max}=6). Besides FSCNet, two multi-path neural networks were designed. These were inspired by the *SHResNet* (Section 5.1.3), however, were trained under the same conditions as the FSCNet to allow a fairer comparison. First and foremost, this meant to train the multi-path neural networks on patches of 11×11×11 voxels rather than 3×3×3 voxels. The success of harmonisation was assessed in a 5-fold cross validation: In five separate training cycles eight subjects were used to learn the model parameters or compute scaling maps and performance of the algorithms was validated on the remaining two subjects. The split into the different training and validation sets was fixed for all experiments. Bootstrapping was not applied, since neural networks take very long to train and the dataset size was very limited.

Linear RISH. After converting DWI data to SH representations in native space, RISH features were computed as described previously (Equation 5.2). Together with b_0 images the four derived RISH feature maps (for each order l=0,2,4,6) were used to calculate a study specific template from all ten available subjects via multi-modal non-linear registration (antsMultivariateTemplateConstruction2.sh as recommended by [126]). For each of the five folds scaling maps were calculated for all image contrasts (b_0 and RISH features) from the respective eight training subjects (Equation 5.8). These scaling maps were then backprojected from template space to native image space for the two remaining subjects. Eventually, SH coefficients were scaled accordingly (Equation 5.7).

CNN Baseline. The baseline model aimed to follow mostly the specifications of the previously described *FSCNet*. It operates on three dimensional patches of 11×11×11 voxels from all SH coefficient maps and the b_0 image. For example, fitting SH up to the 6^{th} order results in 28 SH coefficient maps (l=0: 1 map, l=2: 5 maps, l=4: 9 maps, l=6: 13 maps). Hence, the input would be comprised of 29 images (b_0 and SH images). Input images were all scaled by estimating mean and standard deviation for each volume individually (e.g. b_0 or first SH coefficient map), but collectively for all training scans from one scanner (i.e. Prisma or Connectom). Input patches were passed through four convolutional layers with 75, 150, 300 and 200 feature maps (Figure 5.1). At each layer an isotropic convolutional kernel (3×3×3 voxels) and a dropout rate of 50% was applied. The output of the last convolutional layer was concatenated with the central part (3×3×3 voxels) of the input patch, before eventually feeding it to a bottleneck convolutional layer to reduce the number of output feature maps (Figure 5.1 CONCAT) to the number of b_0 and SH images. The network was then trained on the MUSHAC database sampling 50,000 random patches evenly from all eight training subjects (6250 patches per subject) for each of the 200 epochs. The network parameters were learned on batches of size 20 via Adam optimiser [132] with a learning rate of 10^{-4}. The mean absolute error between output patches (projected Prisma images) and the matching, scaled patches of the reference scanner (Connectom images) served as cost function. Since the CNN learned a mapping between scaled image intensities, the final prediction of the images was generated by backscaling images to the Connectom image space (i.e multiply by standard deviation and adding the mean as estimated from Connectom training scans),

CNN Multi-Path. This network follows the same training scheme as the baseline model, however, instead of processing b_0 and all SH images together, the input was split according to the SH order (Figure 5.1). The b_0 image and the SH coefficient map of order zero were processed together, since they show generally similar anatomical structures. Each of the split input was passed through a block of four convolutional layers (*conv block*). The number of feature maps was chosen to overall match the ones in the baseline model to allow the comparison of networks with similar learning capacity (for example L1: 4×20=80≈75, L3: 4×75=300). A second version of the multi-path CNN was constructed, which averages the output of the *conv blocks* for the higher order SH coefficient maps (i.e. l=4, 6) before adding this value to the input SH coefficients. This aimed to simulate a global scaling of the coefficients rather than learning a local mapping.

Global Scaling. Since the neural networks also included global scaling of SH intensities, this was also implemented as a stand-alone harmonisation approach. In practice, mean and standard deviations for all training input volumes were estimated for Prisma and Con-

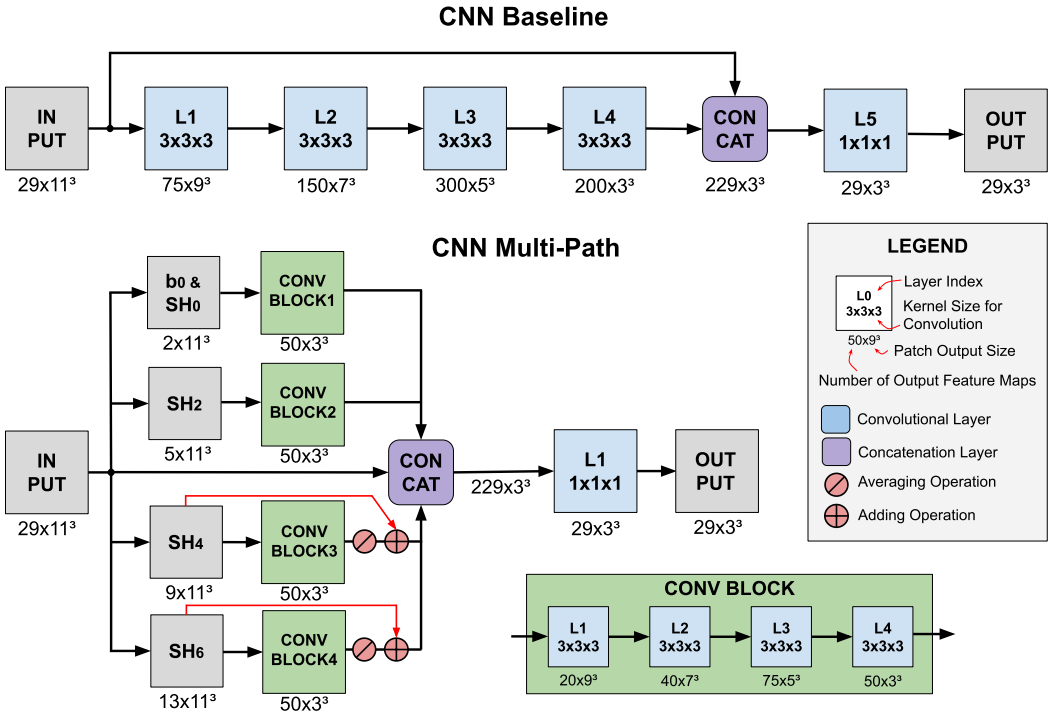


Figure 5.1: Schematic Architecture of Convolutional Neural Networks. The baseline CNN consisted of four convolutional layers that act sequentially on an input of the coefficient images of the SH representations (here l_{max}=6, hence 29 input channels). The output of the last convolutional layer (L4) was concatenated with the input, before going through a bottleneck convolution. The multi-path CNN splits up the SH input images according to their order (l=0,2,4,6) and processes them with four individual convolutional layers (*conv block*). Eventually the different paths are concatenate all together with the input, before feeding them again through a bottleneck convolution. A second approach for the multi-path CNN first averages (*averaging operation*) the output of paths for the higher order SH maps (l=4,6), then adds this average from the SH input images (*adding operation*).

nectom scans separately. Prisma images were mapped to Connectom intensity space by first subtracting the Prisma mean and dividing by the Prisma standard deviation. Then, these scaled images were scaled to Connectom space by multiplying with the Connectom standard deviation and adding the Connectom mean.

Quantitative Assessment. The harmonisation performance of the different models was measured via a global and regional *root mean square error* (RMSE) between the harmonised Prisma RISH feature maps and the Connectom RISH feature maps (analogously for b_0 images):

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_i - y_i)^2} \tag{5.12}$$

with x and y representing the images to be compared and N the number of voxels within a given ROI. The global RMSE was computed within the whole brain mask. Regional

RMSEs were calculated in selected TractSeg ROIs, such as the IFO fascicles (ROI #24 & #25), the ILFs (ROI #26 & #27), the $SLFs_I$ (ROI #35 & #36), the UFs (ROI #43 & #44) and the whole CC (ROI #45).[7] These regions were selected as they were found to be affected in TBI patients (Chapter 3), but also to provide a general overview, rather than comparing all 72 TractSeg ROIs. The underlying assumption was that a lower RMSE reflects less diverging SH coefficient between scanners, which would lead to better harmonised DTI parameter maps. Differences between methods were statistically assessed via rm-ANOVA and post-hoc pairwise Tukey test (GraphPad Prism 8).

5.2.3 Inter-Scanner Variation for CENTER-TBI Substudy

Seven healthy controls, who underwent DWI on both a Trio and Prisma scanner for the CENTER-TBI study, were included. After pre-processing the DWI scans with the diffusion pipeline introduced in section 2.4, images were converted to SH representation ($l_{max} = 4$) and RISH features were calculated. These were used alongside b_0 volumes to create a study specific template via multi-modal registration (antsMultivariateTemplateConstruction2.sh). Subsequently, warped RISH feature images were used to compute CV maps for both scanners separately as well as both datasets combined. The mean and standard deviation was derived on a voxel-wise level. The CV maps were then calculated by dividing the standard deviations by the means again on a voxel-wise level (see Equation 4.2). Mean and standard deviation maps were either computed from Trio scans or Prisma scans only (intra-scanner CV: each seven scans, seven subjects) or from all available scans (inter-scanner CV: 14 scans, seven subjects). For quantitative comparison, mean of CV maps were computed within the whole brain mask and selected TractSeg ROIs (see above).

5.2.4 Denosing of RISH Feature Scaling Maps

Spatial normalisation of brain scans aims to create a voxel-wise correspondence between multiple scans usually from different subjects. The accuracy of this generally depends on the natural anatomical variation across subjects and the success of the deformable registration. To relax the assumption of voxel-wise comparability, images can be smoothed to include for each voxel information from its neighbouring voxels. This can have a denoising effect if voxels with outlier intensities are present. Therefore, the impact of smoothing RISH feature scaling maps for diffusion data harmonisation was examined. For this, the same seven CENTER-TBI controls were used as above. To assess the impact of different scale maps a leave-one-out cross validation was performed. In other words, the corresponding scans of

[7] bilateral ROIs were combined

six subjects were used to compute scaling maps, to then project the RISH feature maps of the remaining validation subject. Smoothing of scaling maps was conducted with either a median filter (kernel size 3) or a Gaussian filter ($\sigma = 1.5$). The b_0 images and RISH feature maps were directly scaled within template space without conversion back to SH space (this also meant not to take the square root of the computed scalemaps). The RMSE error was computed between harmonised Prisma and Trio RISH feature maps. Differences between methods were quantified via rm-ANOVA and post-hoc paired Tukey test.

5.2.5 Subject Selection for RISH Feature Scaling Maps

The initial (standard) approach of RISH feature scaling was later on improved by computing expected RISH feature values only on the basis of the three most similar subjects within the training dataset [187]. The applicability of adaptive subject selection to compute scaling maps was explored for the CENTER-TBI database in three different ways:

1. Computing of scaling maps from the three most similar subjects. For this, the similarity between a test subject's Prisma image and all Prisma training images was estimated via global *mean square error* (MSE, i.e. computing the voxel-wise square error between images and average values within brain mask). The three scans with the lowest MSE and the corresponding Trio scans were used to compute the scaling map.

2. Calculating a weighted average of all training subjects. Scale maps were computed from the weighed mean of the different training subjects. The weighting was defined by the test scans similarity to the training samples (reciprocal value of the relative MSE normalised to the maximum MSE between scans). This aimed to be more adaptive than the standard approach, but more inclusive than the hard selection of three subjects.

3. Computing a weighted average of all training subjects on a voxel-wise basis. Instead of estimating a global weight for the whole feature map, a voxel-wise weighting map was computed. This was based on the reciprocal, relative square error between test and train scans at each voxel. The intention of this approach was to compute more local scaling factors than the other two options.

Scaling maps were all denoised with a median filter (kernel size 3×3×3). Analogous to the previous experiment, the performance was assessed in a leave-one-out cross-validation and by computing the global RMSEs between Trio and scaled Prisma RISH features and b_0 images. Again, differences between approaches were quantified via rm-ANOVA and post-hoc paired Tukey test.

5.2.6 Evaluation of Harmonisation of CENTER-TBI SH Images

Evaluating harmonisation of RISH features in template space can give an indication of a method's performance, however, it is also important to understand the impact on scans in native space. For this, control subjects images on the Prisma scanner were harmonised and compared to the corresponding Trio scans. To allow a voxel-wise comparison between Prisma and Trio scans for each subject, Trio scans were aligned to their corresponding Prisma scan by affinely coregistering b_0 images. This allowed to examine the harmonised Prisma scans and the original Trio scans both in their native space. If Prisma scans were harmonised in a coregistered state, they would need to be backprojected to be analysed in their native subject space. This would lead to multiple interpolation steps and with it blurring images unnecessarily.

Prisma scans were harmonised via linear RISH feature scaling, a CNN with the *Multi-Path* architecture (Section 5.2.2) and global scaling of b_0 and SH volumes. All three methods were evaluated in leave-one-out cross-validation as before.

1. For linear RISH feature scaling, the spatial normalised images of six control subjects[8] (Section 5.2.3) were used to calculate scaling maps for the b_0 and RISH feature images (Equation 5.8). These scaling maps were denoised with a median filter (kernel size 3×3×3) and then backwarped to the native space of the held out test subject scanned on Prisma. Eventually, Prisma SH were scaled as described previously (5.1.3).

2. The hyper-parameters for the *CNN Multi-Path* were set as for previous experiments on the MUSHAC database, with the exception of an increased number of samples extracted per subject during each training epoch. Since only six (instead of eight) subjects were available for training, the samples per subject per epoch were set to 8,320 to keep the total sample number per epoch close to 50,000. Furthermore, SH coefficient could only be estimated up to the 4^{th} order, which is why there were only three convolutional pathways.

3. For comparison, a global scaling of Prisma scans was applied as well. Mean and standard deviation for both scanners were estimated globally for each b_0 image and SH coefficient maps. Each Prisma b_0 and SH volume was then projected to match the mean and variance of Trio scans (see global scaling in Section 5.2.2).

After harmonising SH coefficient images, the RMSE between the RISH feature maps were computed for each subject. Differences between methods were quantified via rm-ANOVA and post-hoc paired Tukey test.

[8] Note: Since multi-modal spatial normalisation is computationally expensive the same template generated for all seven control subjects was used.

5.2.7 Data Harmonisation of CENTER-TBI DTI Metrics

Global and Regional RMSE. To compare the impact of different methods on DTI metrics, the harmonised Prisma SH data were converted back to DWI representation. In addition, the unharmonised Prisma SH data were also backprojected to DWI space to examine the information loss through conversion between representation spaces. After this, FA and MD maps were computed (FSL dtifit with weighted least squares) for all harmonised, backprojected and original Prisma data as well as coregistered Trio data (b-vectors had been rotated accordingly). RMSE was computed both in the whole brain as well as selected ROIs (Section 5.2.2). Differences between methods for the global RMSE were quantified via rm-ANOVA and post-hoc paired Tukey test.

ROI Mean Analysis. Eventually, mean FA and MD values were extracted from the selected TractSeg ROIs (Section 5.2.2) for Prisma scans, native as well as coregistered Trio scans, and harmonised Prisma scans (CNN and Linear-RISH). Besides this, regional mean Prisma values were projected to the Trio domain through ROI-wise standard scaling (analogous to global scaling before, but mean and standard deviation were estimated on a regional level).

5.2.8 Impact of Data Harmonisation on Mild TBI

Another multi-pathway CNN was trained on all seven matched CENTER-TBI control subjects (training parameters the same as for the cross-validation above). This was then applied to b_0 and SH coefficient images of the CENTER-TBI mild TBI patients scanned on Prisma (Table 5.1). Harmonised SH images were than converted back to DWI space. A tensor model was fitted (FSL dtifit with weighted least squares) to extract FA and MD maps for original Prisma, harmonised Prisma and Trio scans in native space. Eventually, regional mean FA and MD values were computed within 24 pre-selected TractSeg ROIs relevant for TBI (Section 3.3.2). As before, Prisma mean DTI metrics were also projected to the Trio domain via ROI-wise standard scaling (see *ROI Mean Analysis* above). Depending on their GOSE at six months post-injury, patients were dichotomised into a group with good (GOSE=8) and poor (GOSE<8) outcome. Then a linear regression analysis (analogous to *Model #3* in Section 3.3.6) was performed to test whether ROI mean DTI metrics were predictive for patient outcome.

5.3 Results

5.3.1 Comparison of Selected Harmonisation Methods

When compared to non-harmonised data, the different approaches for RISH feature harmonisation all led to a reduced global RMSE between matched Prisma and Trio scans (Figure 5.2, repeated measurement ANOVA: $p \leq 0.022$). While linear and global scaling were consistent in performance, CNN approaches produced outliers for b_0 and RISH_0 feature images that had a worse global RMSE than the original inter-scanner difference. This higher variation in harmonisation results could also be observed quantitatively (see standard deviation in Table 5.2). The lowest average global RMSE for RISH_0 features were achieved with linear scaling of RISH feature maps (Linear-RISH). The multi-path CNN, however, achieved lowest global RMSE for b_0 and RISH_2 features. Interestingly, the more local operating method of linear RISH feature scaling performed worse than global scaling of higher order of RISH features (RISH_4: $p_{tukey} = 0.011$; RISH_6: $p_{tukey} = 0.008$). Similarly, global scaling of RISH_6 feature maps reduced RMSE between scanners significantly more than the baseline CNN model (RISH_6: $p_{tukey} = 0.018$). The multi-path CNN approaches could significantly decrease the RMSE between higher order RISH features compared to the baseline CNN (Multi-Path RISH_4 & RISH_6: $p_{tukey} \leq 0.001$; Multi-Path #2 RISH_6: $p_{tukey} = 0.041$). The second approach of the multi-path neural network (*CNN Multi-Path #2*) aimed to emulate globally scaling of RISH_4 and RISH_6 features maps. Indeed, average RMSE were similar to that of global RISH feature map scaling, albeit not statistically significant (Table 5.2 and Figure 5.2).

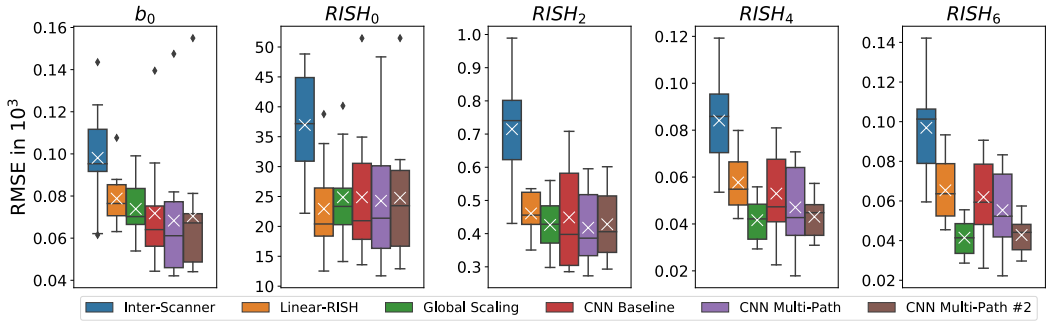


Figure 5.2: Comparison of Global RISH Feature Differences after MUSHAC Data Harmonisation. All five harmonisation methods led to reduced RMSE between. The CNN approaches seemed most successful for b_0 images, however, showed also outliers with higher RMSE than before harmonisation (*Inter-Scanner*). Global scaling was most beneficial for higher order RISH features (RISH_4 and RISH_6). While the multi-path approach (*CNN Multi-Path*) could improve RISH feature harmonisation, only the adjusted second model (*CNN Multi-Path #2*) achieved similarly low levels for higher order RISH features than global scaling.

Although differences across regions could be observed, the general magnitude of inter-scanner

variation before and after harmonisation was similar for all selected ROIs. Harmonisation based on CNNs showed outliers for b_0 images and RISH_0 feature maps with higher RMSE within the examined ROIs than non-harmonised data. Overall, all methods showed similar trends of reduced RMSE for b_0, RISH_0 and RISH_2 images. Stronger differences were found for RISH_4 and RISH_6 feature maps, where the linear RISH feature scaling and the baseline CNN performed worse than the other harmonisation methods. Only the *CNN Multi-Path* networks could reach similar levels for higher order RISH feature maps (l=4,6) than simple global scaling. Although data harmonisation seemed to be slightly more beneficial for UF (e.g RISH_2 or RISH_4), the performances were generally robust across all ROIs (Figure 5.3).

5.3.2 Variation in CENTER-TBI Substudy

Figure 5.4 displays the average of all 14 scans from seven subjects scanned on both scanners and corresponding CV maps. Variation was generally higher in cortical areas where anatomical differences between subjects are more likely. Higher order RISH features (RISH_2 or RISH_4) also showed more variation than b_0 images or zeroth order RISH features (RISH_0). While structured variation was observed in b_0 CV maps, for example around ventricles, RISH_4 feature maps revealed the least structured deviation. Coefficients of variation are similar for scans collected on Prisma or Trio, but increased when combining both datasets. The visual observation of increased variation was also confirmed quantitatively (Table 5.3). Measuring mean CV within the whole brain and selected ROIs, revealed higher variation in all imaged for the pooled dataset in comparison to both individual datasets. Overall, variation in b_0 and RISH_0 feature maps was higher for Trio than for Prisma scans. In contrast, RISH_2 and RISH_4 feature maps mostly demonstrated higher mean CV for Prisma than for Trio scans.

5.3.3 Impact of Denoising on RISH Feature Scaling Maps

Linear scaling of images generally had a positive impact on minimising differences between RISH features from both scanners (Figure 5.5, Table 5.4, all rm-ANOVA: $p \leq 0.023$). However, the standard linear scaling (*Baseline*) significantly increased the RMSE between Trio and Prisma scans for RISH_4 feature maps (Inter-Scanner vs. Baseline: $p_{tukey} = 0.008$). Denoising scaling maps prior to their application could counteract that adverse effect, but not improve RMSE beyond the original inter-scanner differences (Inter-Scanner vs. Median: $p_{tukey} = 0.008$; Inter-Scanner vs. Gaussian: $p_{tukey} = 0.008$). The RMSE between b_0 images and all RISH feature maps could be reduced by denoising scaling maps with a median filter

Table 5.2: Global RMSE for Different Harmonisation Methods. RMSE displayed in 10^3 as mean (median) ± std. Lowest RMSE values for each contrasts across methods are printed in bold.

	Inter-Scanner	Linear-RISH	Global Scaling
b_0	0.098 (0.095) ± 0.025	0.079 (0.076) ± 0.013	0.074 (0.070) ± 0.014
$RISH_0$	36.97 (37.18) ± 9.50	**22.90** (**20.40**) ± 8.20	24.87 (23.34) ± 7.82
$RISH_2$	0.715 (0.741) ± 0.177	0.461 (0.456) ± 0.067	0.426 (0.428) ± 0.080
$RISH_4$	0.084 (0.086) ± 0.019	0.058 (0.055) ± 0.013	**0.042** (**0.042**) ± 0.009
$RISH_6$	0.097 (0.101) ± 0.024	0.065 (0.064) ± 0.016	**0.041** (**0.041**) ± 0.009

	CNN Baseline	CNN Multi-Path	CNN Multi-Path #2
b_0	0.072 (0.064) ± 0.028	**0.068** (**0.061**) ± 0.031	0.070 (0.067) ± 0.032
$RISH_0$	24.87 (20.95) ± 11.79	24.28 (21.37) ± 11.45	24.79 (23.45) ± 11.44
$RISH_2$	0.449 (0.398) ± 0.159	**0.418** (**0.386**) ± 0.115	0.428 (0.406) ± 0.109
$RISH_4$	0.053 (0.047) ± 0.019	0.047 (0.043) ± 0.018	0.043 (0.045) ± 0.009
$RISH_6$	0.062 (0.059) ± 0.021	0.055 (0.052) ± 0.020	0.043 (0.044) ± 0.009

P-Values of Post-Hoc Paired Tukey-Test						
Method #1	Method #2	b_0	$RISH_0$	$RISH_2$	$RISH_4$	$RISH_6$
Inter-Scanner	Linear-RISH	0.271	0.116	**0.012**	**<.001**	**<.001**
Inter-Scanner	Global Scaling	0.177	0.189	**0.017**	**<.001**	**<.001**
Inter-Scanner	CNN Baseline	0.206	0.318	0.107	0.056	0.056
Inter-Scanner	CNN Multi-Path	0.148	0.264	**0.030**	**0.021**	**0.026**
Inter-Scanner	CNN Multi-Path #2	0.159	0.273	**0.026**	**<.001**	**<.001**
Linear-RISH	Global Scaling	**0.016**	**0.029**	0.251	**0.011**	**0.008**
Linear-RISH	CNN Baseline	0.753	0.979	1.000	0.9828	0.998
Linear-RISH	CNN Multi-Path	0.564	0.993	0.578	0.668	0.840
Linear-RISH	CNN Multi-Path #2	0.803	0.988	0.725	**0.007**	**0.006**
Global Scaling	CNN Baseline	0.998	>0.999	0.985	0.197	**0.018**
Global Scaling	CNN Multi-Path	0.995	>0.999	0.998	0.764	0.099
Global Scaling	CNN Multi-Path #2	0.995	>0.999	>0.999	0.207	0.226
CNN Baseline	CNN Multi-Path	0.633	0.947	0.693	**<.001**	**<.001**
CNN Baseline	CNN Multi-Path #2	0.994	>0.999	0.906	0.395	**0.041**
CNN Multi-Path	CNN Multi-Path #2	0.783	0.988	0.806	0.944	0.214
rm-ANOVA p-val		**0.022**	**<.001**	**0.003**	**0.001**	**<.001**

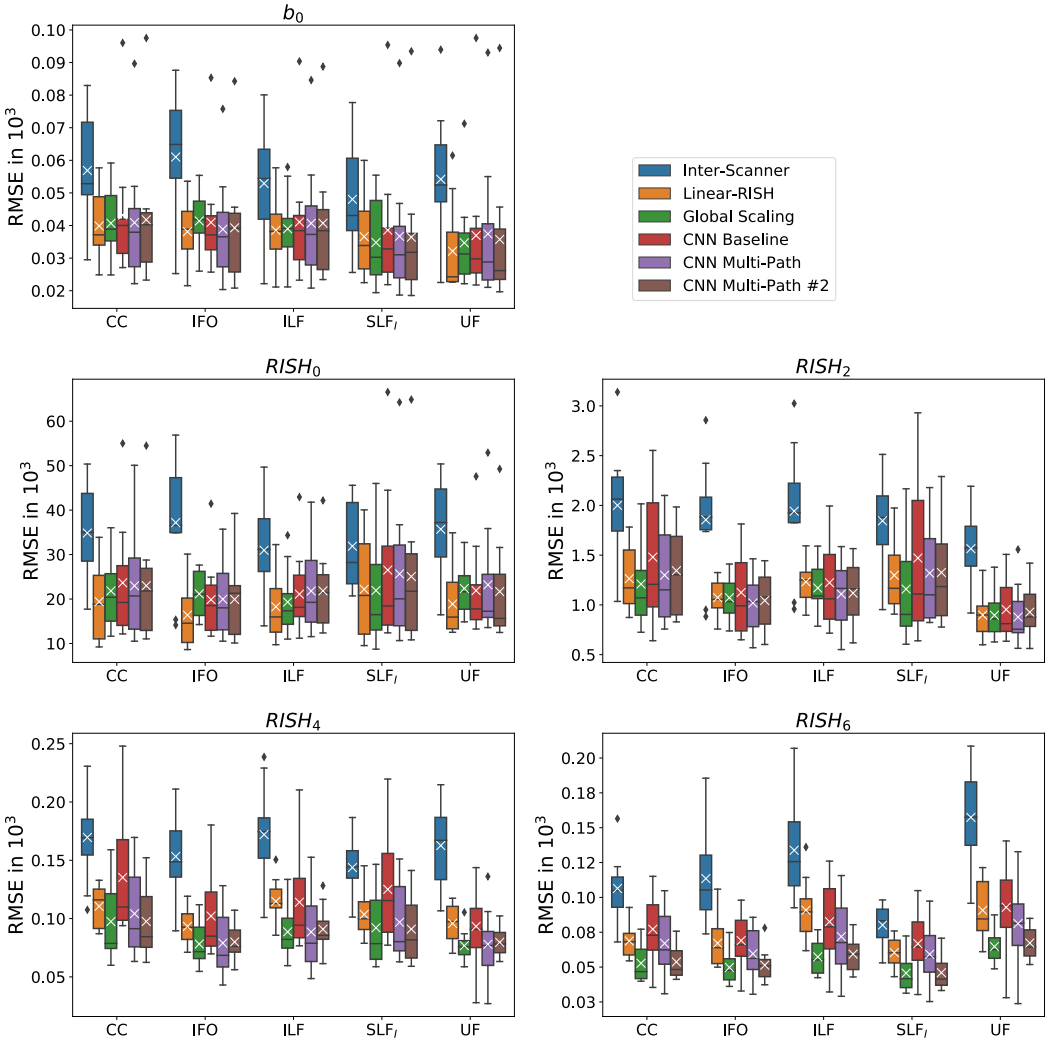


Figure 5.3: Comparison of Regional RISH Feature Differences after MUSHAC Data Harmonisation. The extend of inter-scanner differences varied for different regions and different image contrasts (b_0 and RISH features. All harmonisation approaches helped to reduced the RMSE between scanners, however, CNN approaches resulted in outlier cases with higher RMSE than before harmonisation (particularly for b_0 and RISH_0). Overall, global scaling and the CNN multi-path approaches were more successful to harmonise RISH_4 features. The CNN baseline showed a high variation is RMSE for most higher order RISH features (l=2,4,6). Linear RISH features scaling seemed most beneficial for RISH_0 features.

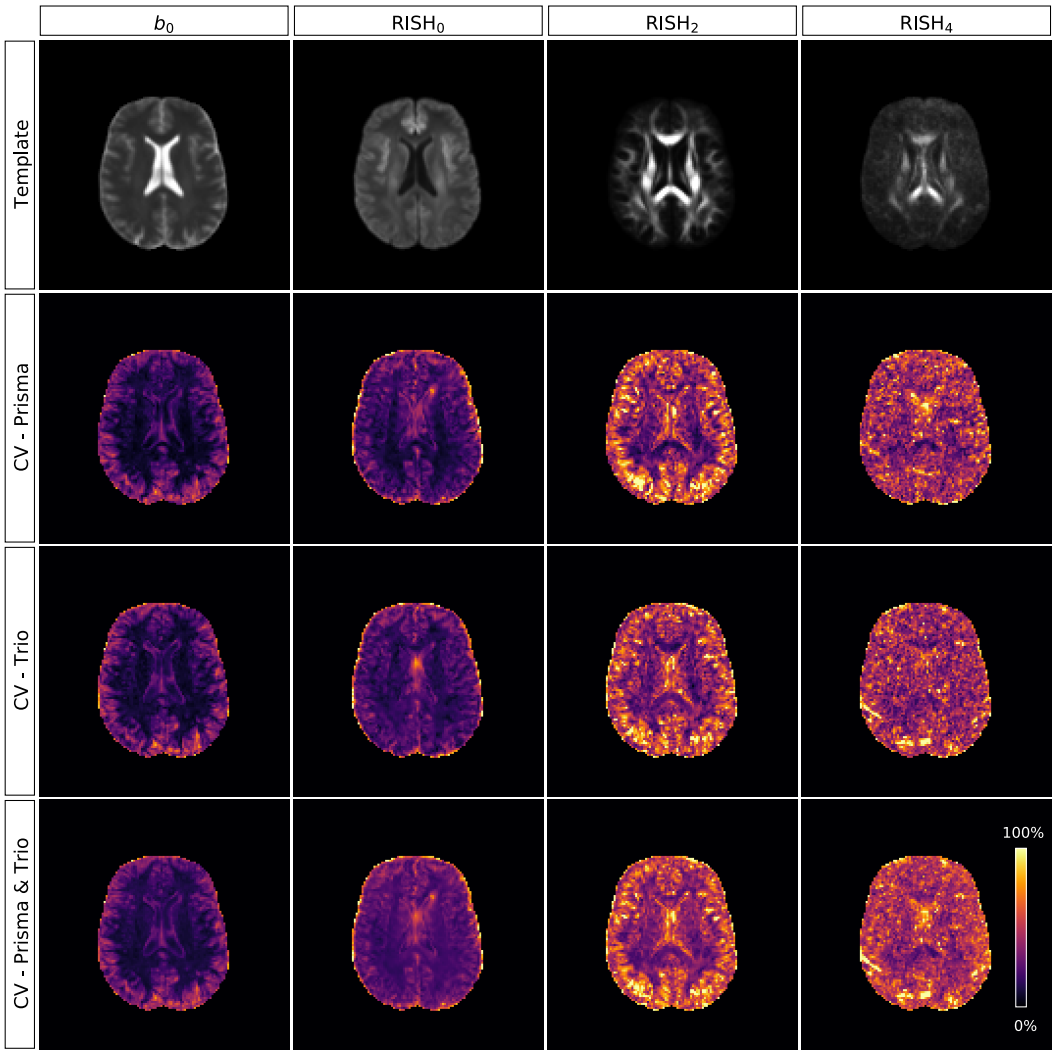


Figure 5.4: Intra- and Inter-Scanner Variation for CENTER-TBI Cambridge Subset. **Top row:** Average b_0 images and RISH feature maps for all 14 scans from seven subjects scanned on Prisma and Trio. **Row 2-4:** Lowest variation within and across scanners was found in the WM of b_0 images. Cortical areas showed generally higher variance, particularly on RISH$_2$ feature maps. Variation was similar within both scanners (both middle rows), but increased when combining data from both scanners (bottom row). All CV maps are shown using the same scale.

Table 5.3: Global and Regional Variation within and Across Scanners for CENTER-TBI Subset. Coefficients of variation displayed as mean ± std in %. Mean values larger than the equivalent metric at the other single scanner data printed in bold.

	ROI	b_0	RISH_0	RISH_2	RISH_4
Prisma	Whole Brain	23.2 ± 15.2	28.1 ± 20.0	**52.0** ± 25.2	**48.6** ± 19.8
	CC	16.3 ± 12.4	17.5 ± 8.5	**45.6** ± 23.4	41.6 ± 15.4
	IFO	16.2 ± 11.0	17.7 ± 8.9	**47.2** ± 23.7	**45.2** ± 17.5
	ILF	12.9 ± 7.6	17.7 ± 9.3	**44.5** ± 20.5	**46.4** ± 19.0
	SLF_I	12.5 ± 11.5	13.6 ± 5.6	39.1 ± 19.9	37.3 ± 13.8
	UF	12.5 ± 6.5	**23.5** ± 11.3	**38.1** ± 16.5	**46.9** ± 17.1
Trio	Whole Brain	**24.0** ± 14.8	**28.2** ± 17.2	50.2 ± 21.5	48.3 ± 19.6
	CC	**17.5** ± 12.1	**19.1** ± 7.4	45.0 ± 19.7	**42.2** ± 15.2
	IFO	**16.7** ± 10.5	**19.3** ± 7.1	45.2 ± 20.1	43.8 ± 17.4
	ILF	**14.2** ± 7.5	**20.5** ± 8.0	43.9 ± 17.9	45.4 ± 18.2
	SLF_I	**13.8** ± 11.2	**15.9** ± 5.5	**42.1** ± 16.9	**40.1** ± 13.2
	UF	12.5 ± 6.6	20.6 ± 9.4	36.0 ± 13.0	43.1 ± 15.0
Prisma & Trio	Whole Brain	24.7 ± 13.6	32.5 ± 17.4	53.8 ± 21.8	55.1 ± 19.4
	CC	18.4 ± 11.1	23.2 ± 6.6	47.0 ± 19.3	47.4 ± 15.0
	IFO	18.4 ± 9.7	24.2 ± 6.9	48.4 ± 20.2	48.9 ± 16.8
	ILF	15.1 ± 6.5	22.8 ± 6.7	45.5 ± 17.5	51.2 ± 18.0
	SLF_I	14.7 ± 10.3	20.0 ± 4.1	42.4 ± 16.1	44.1 ± 12.9
	UF	14.8 ± 5.3	26.4 ± 9.1	41.0 ± 13.8	48.8 ± 14.2

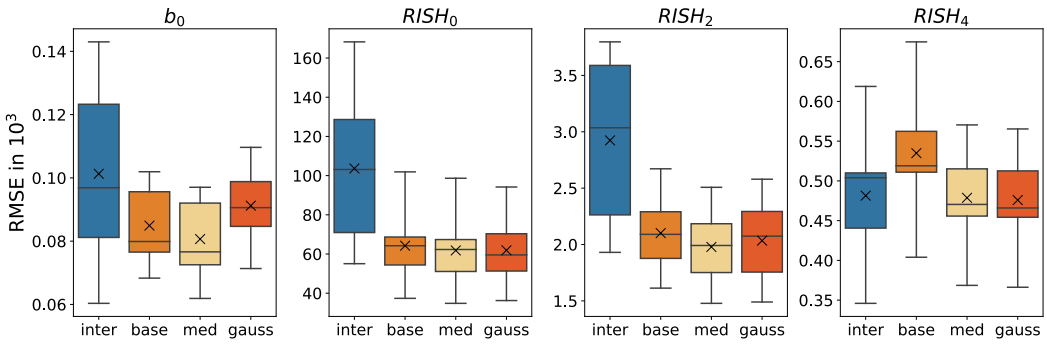


Figure 5.5: Effect of Denoising of Scaling Maps for Data Harmonisation. Linear scaling of b_0 images and RISH features had a positive effect on reducing the RMSE (base) in comparison to the inter-scanner differences (inter). The exception was RISH_4 features for which an increase in RMSE was observed. Denoising scaling maps with a median filter (med) had overall a small but beneficial effect on further reducing the RMSE. Gaussian filtering (gauss) was less effective and negatively impacted b_0 image scaling.

(Baseline vs. Median: all $p_{tukey} \leq 0.006$). This was not observed for Gaussian filtering, with the exception of RISH_4 features maps (Baseline vs. Gaussian: $p_{tukey} \leq 0.036$).

Table 5.4: Impact of Denoising of Scaling Maps on Harmonisation between Scanners for CENTER-TBI Subset. RMSE displayed as mean ± std $\times 10^3$. P-values<0.05 printed in bold (no correction for multiple comparison).

Method	b_0	RISH_0	RISH_2	RISH_4
Inter-Scanner	0.10 ± 0.03	103.65 ± 40.80	2.92 ± 0.77	0.48 ± 0.08
Baseline	0.08 ± 0.01	64.23 ± 20.60	2.10 ± 0.36	0.53 ± 0.08
Median	0.08 ± 0.01	61.81 ± 20.69	1.98 ± 0.35	0.48 ± 0.06
Gaussian	0.09 ± 0.01	61.89 ± 19.32	2.03 ± 0.39	0.48 ± 0.06

P-Values of Post-Hoc Paired Tukey-Test					
Method #1	Method #2	b_0	RISH_0	RISH_2	RISH_4
Inter-Scanner	Baseline	0.335	0.082	0.069	**0.008**
Inter-Scanner	Median	0.231	0.073	**0.043**	0.994
Inter-Scanner	Gaussian	0.869	0.093	**0.024**	0.959
Baseline	Median	**0.005**	**0.003**	**0.002**	**0.006**
Baseline	Gaussian	0.569	0.389	0.569	**0.005**
Median	Gaussian	0.151	1.000	0.709	**0.036**
rm-ANOVA p-val		0.161	**0.023**	**0.012**	**<0.001**

5.3.4 Scan Selection and Weighting for Scaling Maps

Scaling Maps. Figure 5.6 shows an example of differently computed scaling maps for one healthy CENTER-TBI subject. The baseline scaling maps, computed from all available training subjects, showed the most spatially continuous values. Selecting only the three most similar subjects to compute scaling maps resulted in higher variation. Weighting the subjects according to image similarity resulted in very similar scaling maps that obtained from the baseline approach (basically, weighting all subject equally). Computing scaling values for each voxel independently resulted in more speckled scaling maps. For all approaches the variation in scaling values was highest for RISH_4 features. Scaling maps for b_0 images showed the most values close to one, which compounds to no change between scanner intensities.

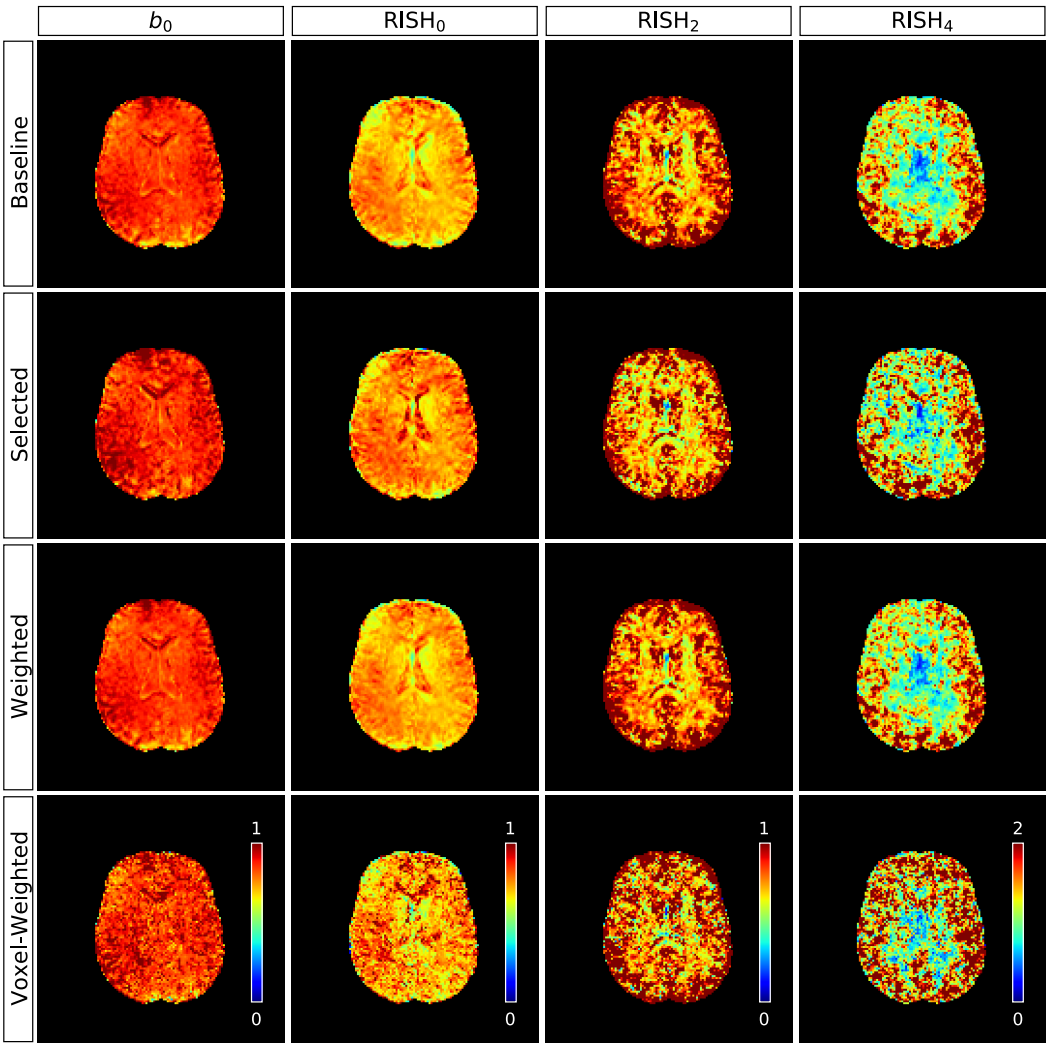


Figure 5.6: Example of Scaling Maps of CENTER-TBI Subject. Selecting a subset of subjects resulted in slightly more structured scaling maps (*Selected*). Weighting the subjects (*Weighted*) obtained very similar scaling maps compared to when no weighting was applied (*Baseline*). Computing a voxel-wise weighting resulted in spatially less continuous scaling values (*Voxel-Weighted*). Most variation was found for RISH$_4$ features. The least impact of scaling was observed for b_0 images.

Global RMSE. Overall, differences between the various calculations of scaling maps were subtle (Figure 5.7). Voxel-wise weighting resulted in the lowest RMSE between scanners for all RISH feature maps (Table 5.5). Voxel-wise weighting was also significantly better than global weighting (Weighted vs. Voxel-Weighted: $p_{tukey} = 0.001$). Selecting the three most similar subjects to calculate scaling maps (*Selected*) led to higher RMSEs than any other approach, which was most obvious for RISH_4 feature maps (post-hoc paired Tukey test: all $p_{tukey} \leq 0.025$). No direct impact of different computation methods for scaling maps was observable for b_0 images (post-hoc paired Tukey test: all $p_{tukey} \geq 0.962$).

Table 5.5: Impact of Subject Selection and Weighting on Harmonisation for CENTER-TBI Subset. RMSE displayed as mean ± std $\times 10^3$. P-values<0.05 printed in bold.

Method	b_0	RISH_0	RISH_2	RISH_4
Inter-Scanner	0.10 ± 0.03	103.65 ± 40.80	2.92 ± 0.77	0.48 ± 0.08
Baseline	0.08 ± 0.01	61.81 ± 20.69	1.98 ± 0.35	0.48 ± 0.06
Selected	0.08 ± 0.01	62.87 ± 22.88	2.07 ± 0.38	0.52 ± 0.06
Weighted	0.08 ± 0.01	61.72 ± 20.61	1.98 ± 0.35	0.48 ± 0.06
Voxel-Weighted	0.08 ± 0.01	**60.10** ± 19.20	**1.94** ± 0.31	**0.45** ± 0.05

P-Values of Post-Hoc Paired Tukey-Test					
Method #1	Method #2	b_0	RISH_0	RISH_2	RISH_4
Inter-Scanner	Baseline	0.251	0.098	0.060	1.000
Inter-Scanner	Selected	0.206	0.077	0.085	0.060
Inter-Scanner	Weighted	0.274	0.092	0.060	1.000
Inter-Scanner	Voxel-Weighted	0.311	0.092	0.063	0.240
Baseline	Selected	0.997	0.998	0.128	**0.025**
Baseline	Weighted	0.996	0.976	0.996	1.000
Baseline	Voxel-Weighted	0.998	0.757	0.933	**0.002**
Selected	Weighted	0.994	0.995	0.154	**0.018**
Selected	Voxel-Weighted	1.000	0.989	0.303	**0.003**
Weighted	. Voxel-Weighted	0.962	0.919	0.904	**0.001**
rm-ANOVA p-val		0.061	**0.015**	**0.012**	**<0.001**

Median filtering was applied to all four approaches.

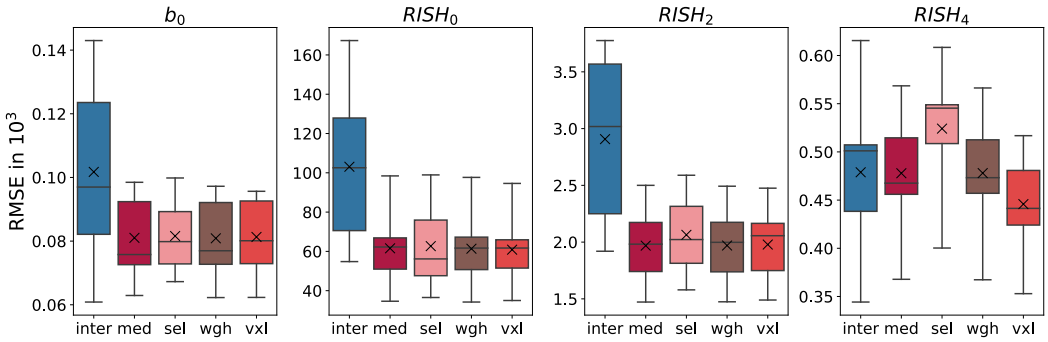


Figure 5.7: Impact of Selection of Subjects on Image Harmonisation. All methods reduced the global RMSE for b_0, RISH_0 and RISH_2, but differences between approaches were marginal. Selecting a subset of subjects (sel) increased the global RMSE for RISH_4. Only voxel-wise weighting (vxl) had an observable positive effect on RISH_4 feature maps. All approaches included the same median filtering of scaling maps.

5.3.5 Harmonisation of CENTER-TBI SH Images

Figure 5.8 shows the impact of different harmonisation methods on the inter-scanner variance (measured via RMSE) for the CENTER-TBI control subjects. Linear RISH feature scaling, global scaling and the CNN resulted in reduced RMSEs between Prisma and Trio scans for b_0 images as well as RISH_0 and RISH_2 features. While the neural network could also successfully harmonise SH coefficients of higher order (RISH_4), both other methods seemed to fail to minimise the RMSE. In fact, global scaling resulted in increased RMSE between both scanners for RISH_4 features maps. Looking at RMSE for individual subjects showed that harmonisation was not entirely consistent for all subjects. For example, two subjects (B & G) had a higher RMSE for RISH_2 features after harmonisation with CNN than with linear scaling (Linear-RISH or Global Scaling).

All three harmonisation methods reduced the RMSE compared to non-harmonised data (Table 5.6). The exception was global scaling for RISH_4 features which yielded significantly higher RMSE (Inter-Scanner vs. Global-Scaling: $p_{tukey} < 0.001$). Linear RISH feature scaling was beneficial to reduce RMSE for RISH_0 and RISH_2 feature maps (Inter-Scanner vs. Linear-RISH: $p_{tukey} = 0.005$ and $p_{tukey} < 0.001$, respectively). The lowest mean RMSE for b_0 images and all RISH feature maps were achieved by the CNN Multi-Path. The neural network outperformed both the linear RISH feature and global scaling for b_0 and RISH_4 images (All: $p_{tukey} \leq 0.025$). Global scaling performed worse on higher order RISH features.

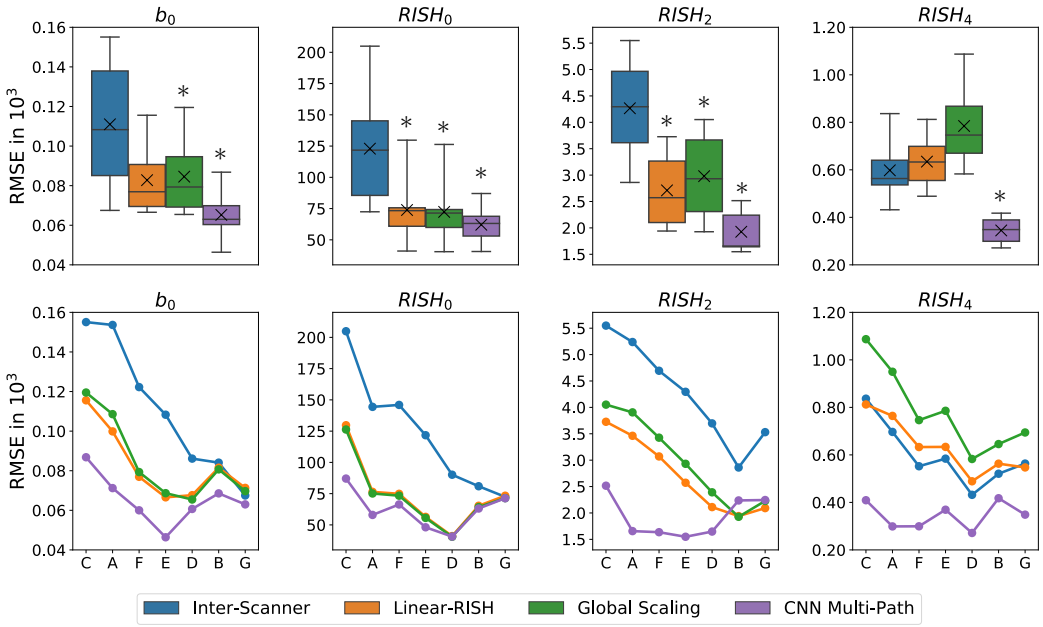


Figure 5.8: Comparison of Harmonisation Methods on CENTER-TBI Controls. **Top:** All three methods were able to reduced the global RMSE between the data from different scanners for b_0 images, RISH_0 and RISH_2 feature maps. Global scaling was overall similar in performance to linear RISH feature scaling, however, worsened the inter-scanner differences for RISH_4 feature maps. The CNN Multi-Path provided the lowest RMSE. Significantly lower RMSE than Inter-Scanner RMSE marked with an asterisk. **Bottom:** RMSE for individual subjects (A-G) sorted according to descending inter-scanner RMSE for b_0 images. Harmonisation for subjects B and G was less successful than for other subjects.

5.3.6 Evaluation of Harmonisation for CENTER-TBI DTI Metrics

Global RMSE. Figure 5.9 shows the global RMSE of FA and MD maps before and after harmonisation. For both DTI metrics, the differences in RMSE were all highly significant between all harmonisation methods as well as no data harmonisation. Backprojection of SH to DWI space led only to marginal, but significant (FA: $p_{tukey} = 0.025$, MD: $p_{tukey} < 0.001$) differences.[9] Linear RISH feature scaling had a negative effect on both DTI metrics as the observed RMSE was significantly increased after harmonisation (both for FA and MD Inter-Scanner vs. Linear-RISH: $p_{tukey} < 0.001$). Both global scaling and CNN Multi-Path harmonisation could significantly reduce the deviation between DTI metrics from the two scanners (All: $p_{tukey} < 0.001$). The CNN Multi-Path harmonisation achieved overall the lowest average of global and regional RMSE among the methods analysed (Table 5.7).

[9]Changes are not directly reflected in the mean RMSE, but were consistent for all subjects; RMSE was lower for FA and higher for MD for all subjects after backprojection.

Table 5.6: Comparison of Harmonisation Methods on Control Subjects from CENTER-TBI. Global RMSE displayed as mean ± std $\times 10^3$. Lowest mean RMSE per contrast printed in bold.

Method	b_0	RISH_0	RISH_2	RISH_4
Inter-Scanner	0.11 ± 0.03	122.93 ± 46.73	4.27 ± 0.97	0.60 ± 0.13
Linear-RISH	0.08 ± 0.02	73.87 ± 27.63	2.71 ± 0.72	0.63 ± 0.12
Global Scaling	0.08 ± 0.02	72.36 ± 26.71	2.98 ± 0.84	0.78 ± 0.18
CNN Multi-Path	**0.07** ± 0.01	**62.08** ± 15.27	**1.93** ± 0.39	**0.34** ± 0.06

P-Values of Post-Hoc Paired Tukey-Test					
Method #1	Method #2	b_0	RISH_0	RISH_2	RISH_4
Inter-Scanner	Linear-RISH	0.059	**0.005**	**<0.001**	0.184
Inter-Scanner	Global Scaling	**0.043**	**0.004**	**<0.001**	**<0.001** *
Inter-Scanner	CNN Multi-Path	**0.026**	**0.008**	**0.005**	**0.006**
Linear-RISH	Global Scaling	0.630	**0.006** +	**0.015** +	**0.004** +
Linear-RISH	CNN Multi-Path	**0.008**	0.082	0.141	**0.002**
Global Scaling	CNN Multi-Path	**0.025**	0.103	0.093	**0.002**
rm-ANOVA p-val		**0.007**	**0.005**	**0.002**	**<0.001**

*global scaling significantly *worse* than Inter-Scanner RMSE, +global scaling significantly *worse* than Linear-RISH RMSE

Regional RMSE. A similar observation was made for regional FA metrics (Figure 5.10). Within the ROIs examined, the RMSE was increased after harmonisation with linear RISH feature scaling. In contrast, both global scaling of FA maps as well as the neural network harmonisation could reduce the RMSE between matched scans from Prisma and Trio. The CNN approach showed the best potential to minimise differences between scans. Nonetheless, harmonisation performance varied for the regions investigated. For example, SLF_I and UF showed low and very high variation of RMSE, respectively (Table 5.7).

Harmonisation of the MD metric between scanners showed a different pattern. Although both linear RISH feature and global scaling could reduce the RMSE for some regions, the effect of data harmonisation was marginal. Globally, the RMSE was increased for MD values after linear RISH feature scaling (Figure 5.9), however, within selected WM tracts, such as the SLF_I and UF, this harmonisation approach could decrease inter-scanner differences. The RMSEs in both ROIs were increased after global scaling of MD metrics. Overall, harmonisation via global scaling worked much better for FA than MD values. The neural network was most successful in lowering regional RMSE between images from both scanners, which was consistent with the observation made for the global RMSE.

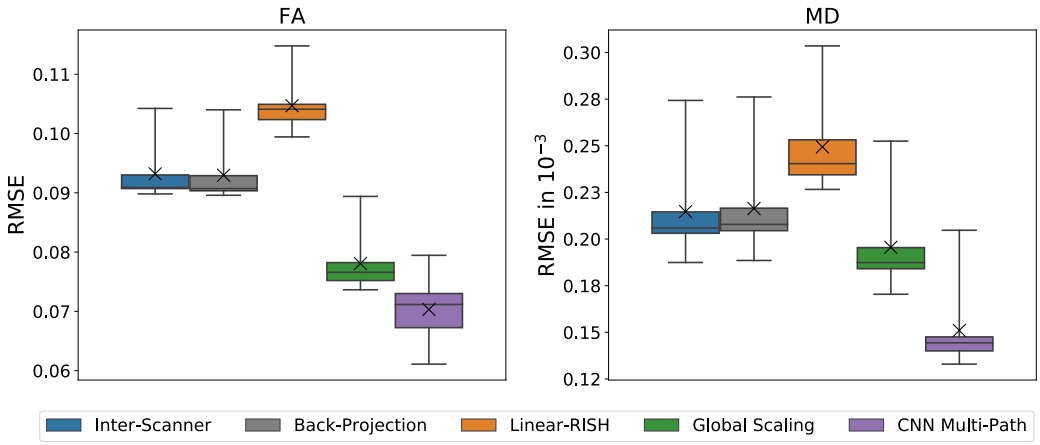


Figure 5.9: Comparison of DTI Metrics Before and After Harmonisation for CENTER-TBI Controls. Backprojecting the data from SH to DWI representation did not show any impact on the RMSE between scans from Prisma and Trio. Linear RISH feature scaling resulted in an increase of global RMSE for both FA and MD. While global scaling could lower the RMSE between scanners, the CNN Multi-Path showed the best potential for harmonising DTI metrics.

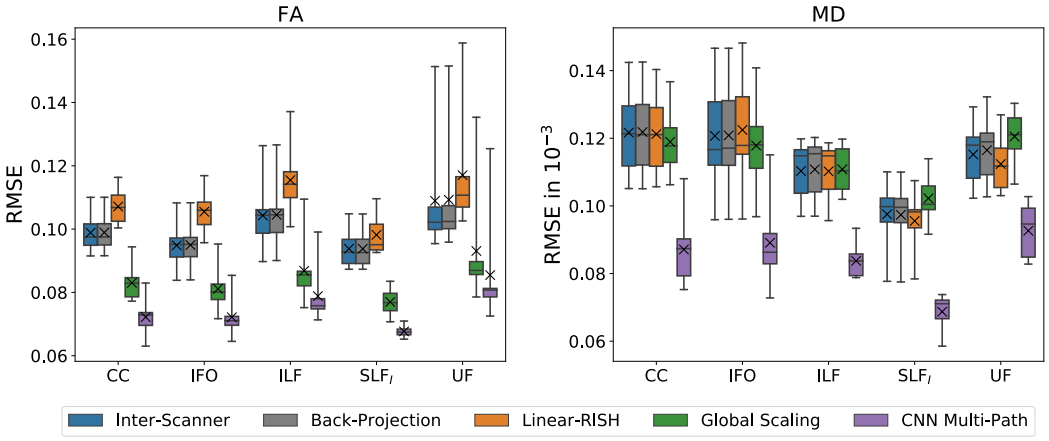


Figure 5.10: Comparison of Regional DTI Metrics Before and After Harmonisation for CENTER-TBI Controls. The RMSE between FA from different scanners could be reduced via global scaling and the CNN Multi-Path harmonisation, whereas the latter seemed more successful. Linear RISH feature scaling increased the RMSE of regional FA. Both linear RISH feature and global scaling showed no clear reduction of RMSE between MD values across scanners. Regions such as the SLF$_I$ and the UF slightly benefited from linear RISH feature harmonisation (reduced RMSE compared to inter-scanner differences). Only the CNN Multi-Path approach displayed reduced RMSE for regional MD between scanners.

Table 5.7: Global and Regional RMSE for FA and MD Within and Across Scanners for CENTER-TBI Controls. All RMSE displayed as mean ± std. MD represented as $\times 10^{-3}$

	ROI	Inter-Scanner	Linear-RISH	Global Scaling	CNN M.P.
FA	Global	0.093 ± 0.005	0.105 ± 0.005	0.078 ± 0.005	0.070 ± 0.006
	CC	0.099 ± 0.006	0.107 ± 0.006	0.083 ± 0.006	0.072 ± 0.006
	IFO	0.095 ± 0.008	0.105 ± 0.007	0.081 ± 0.007	0.072 ± 0.006
	ILF	0.104 ± 0.011	0.115 ± 0.011	0.087 ± 0.011	0.079 ± 0.009
	SLF_I	0.094 ± 0.006	0.098 ± 0.007	0.077 ± 0.004	0.068 ± 0.002
	UF	0.109 ± 0.019	0.117 ± 0.019	0.093 ± 0.019	0.085 ± 0.018
MD	Global	0.215 ± 0.028	0.249 ± 0.026	0.196 ± 0.027	0.151 ± 0.024
	CC	0.122 ± 0.013	0.121 ± 0.013	0.119 ± 0.010	0.087 ± 0.011
	IFO	0.121 ± 0.017	0.122 ± 0.017	0.118 ± 0.014	0.089 ± 0.014
	ILF	0.110 ± 0.009	0.110 ± 0.009	0.111 ± 0.007	0.084 ± 0.005
	SLF_I	0.098 ± 0.010	0.096 ± 0.009	0.102 ± 0.007	0.069 ± 0.005
	UF	0.115 ± 0.009	0.112 ± 0.009	0.121 ± 0.008	0.093 ± 0.008

P-Values of Post-Hoc Paired Tukey-Test			
Method #1	Method #2	FA	MD
Inter-Scanner	Linear-RISH	**<0.001**	**<0.001**
Inter-Scanner	Global Scaling	**<0.001**	**<0.001**
Inter-Scanner	CNN Multi-Path	**<0.001**	**<0.001**
Linear-RISH	Global Scaling	**<0.001**	**<0.001**
Linear-RISH	CNN Multi-Path	**<0.001**	**<0.001**
Global Scaling	CNN Multi-Path	**0.003**	**<0.001**
rm-ANOVA p-val		**<0.001**	**<0.001**

Values for *Backprojection* not displayed for brevity and since they were near equal to Inter-Scanner values (FA: $p_{tukey} = 0.025$; MD: $p_{tukey} < 0.001$). CNN M.P. stands for CNN Multi-Path approach.

Regional Mean DTI Metrics. Comparing mean FA metrics within the selected ROIs revealed only marginal differences between Prisma and Trio scans (Figure 5.11 left). Interestingly, there was an obvious drop in mean FA values of coregistered Trio scans compared to the original scans. The multi-path CNN had been trained to map Prisma scans to the coregistered Trio scans. This was also reflected in the harmonised regional mean FA values (*Prisma - CNN Multi-Path*), as these match the FA values from coregistered Trio scans (*Trio Coreg*) much more closely than the ones extracted from the native scans (*Trio*). Linear RISH and ROI-wise scaling resulted in slightly increased FA values compared to native Trio FA values. The differences between scanners were not as obvious for MD values (Figure 5.11 right). Linear RISH scaling and CNN harmonisation showed increased MD values in the SLF_I, despite Trio MD values being lower than unharmonised Prisma MD values. Mapping Prisma MD values to Trio space via ROI-wise scaling also resulted in outliers (e.g. reduced MD in the CC).

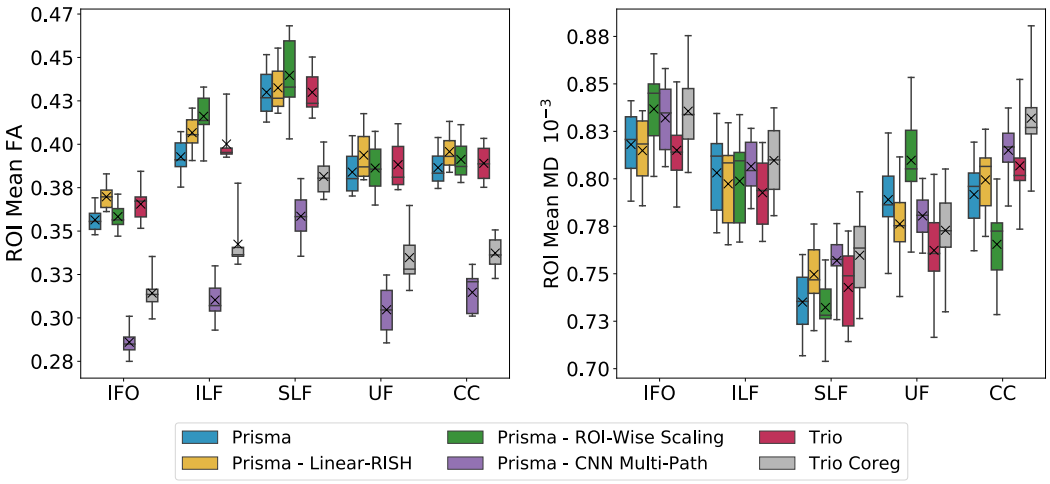


Figure 5.11: Comparison of Means DTI Metrics in Controls Before and After Harmonisation. Regional mean values were marginally different between Trio and Prisma scans. Both Linear-RISH and ROI-wise scaling overly increased Prisma FA values (e.g. ILF). The CNN Multi-Path harmonisation strongly decreased FA values in comparison to the original values, but matched most closely regional mean FA values extracted from coregistered Trio scans. Differences in MD between scanners were not as drastic as for FA. However, ROI-wise scaling led to some outliers (i.e. UF or CC).

5.3.7 Impact of Data Harmonisation on Mild TBI

The regression analysis showed that none of the 24 selected TractSeg regions was predictive of patient outcome. This result was consistent regardless whether the examined scans came from one of the two scanners, both scanners, or the pooled Trio and harmonised Prisma

data. Nonetheless, a general trend of lower FA and higher MD values was observed in TBI patients with poor outcome. As for the control subjects from CENTER-TBI, regional DTI metrics were very similar for both Prisma and Trio scans, and the CNN harmonisation decreased FA metrics substantially. Furthermore, it seemed that the separation between patients with good and poor outcome increased through CNN harmonisation. Regional MD values seemed mostly robust to any applied harmonisation method and strong effects could not be observed visually (Figure 5.12).

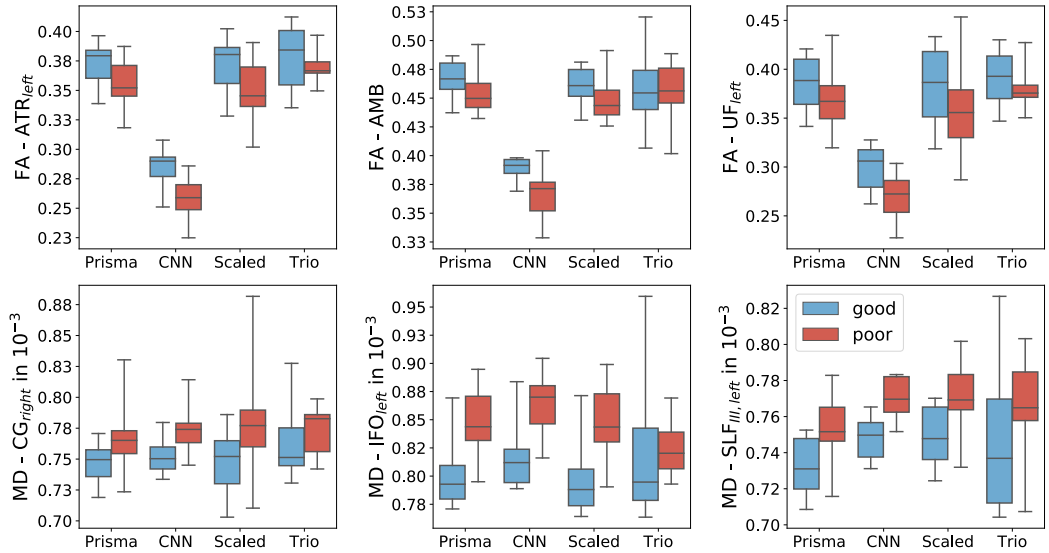


Figure 5.12: Regional Mean FA and MD Before and After Harmonisation for Mild TBI Patients in Selected Regions. Harmonisation via CNN substantially reduced the FA mean values. This could not be observed for MD values, which seemed mostly stable regardless whether data harmonisation was applied or not.

5.4 Discussion

5.4.1 Potential and Limitations of Harmonisation Methods

MUSHAC Database. Although not the strongest harmonisation method, the linear RISH feature scaling performed most consistently in lowering the RMSE between b_0 and RISH feature images of the MUSHAC benchmark database. In contrast, all three neural network approaches showed strong RMSE outliers for b_0 and RISH$_0$ images. Generally, this could indicate an overfitting of the model. However, upon visual inspection, the outlier RMSE metric could be associated with a particular Prisma scan that had a higher intensity range than the other Prisma scans even in unharmonised state. Later on, this particular subject was also excluded from a different experiment conducted by the authors who provided the

MUSHAC database [244]. Nonetheless, this shows how machine learning models can be sensitive to outliers and do not necessarily generalise well. Such a sensitivity could be problematic for analysis of patient data, as a neural network trained on control subjects could artificially enhance abnormalities found in patient scans. Global scaling of SH volumes performed surprisingly well and in particular outperformed more complex methods for RISH_4 and RISH_6 features in the MUSHAC dataset. While b_0 images and lower order SH images display anatomical structures, higher order SH images are more noisy. Approaches that harmonise on a local level (Linear-RISH & CNN Baseline and CNN Multi-Path) may fail to learn to scale unstructured image intensities. Therefore, methods based on global scaling (including CNN Multi-Path #2) were more beneficial for higher order SH harmonisation. Previous publications on the same database [187, 238] have only assessed harmonisation methods based on lower order RISH features (i.e. RISH_0 and RISH_2). The experiments presented in this chapter have shown that a neural network that processes SH of different orders on multiple individual pathways performs better than a neural network with one pathway. This and the fact that global scaling for higher order SH images was more beneficial, is a clear indicator, that information from SH images may need to be harmonised differently dependent on the SH order.

CENTER-TBI Database. When comparing RISH features from the CENTER-TBI database, the multi-path neural network performed best to reduce the RMSE between Trio and Prisma scans. Although a high variation for the different models was observed, Cohen's measure showed that the harmonisation was indeed effective. All three harmonisation approaches could reduce inter-scanner differences for b_0, RISH_0 and RISH_2 images (Cohen's effect size $d > 0.2$). The strongest effect was observed for CNN Multi-Path (Cohen's effect size $d > 1.3$), which also was more effective than Linear-RISH harmonisation (Cohen's effect size $d > 0.5$) and global image scaling (Cohen's effect size $d > 0.4$) for all four images (Table 7.4 in Appendix). The Linear-RISH harmonisation failed to harmonise SH of the 4^{th} order (RISH_4) and could not significantly reduce differences between b_0 images. This could be a reason why RMSEs between FA and MD maps were increased after Linear-RISH harmonisation. Nonetheless, this could also be a consequence of previous processing steps: The RMSE computed for validation is based on averaging on voxel-wise differences between the harmonised images and the coregistered Trio scans. The CNN was trained on the matching pairs of Prisma and coregistered Trio scans, hence could learn a direct mapping to the coregistered Trio scans, later also used for validation. In contrast, the linear RISH feature scaling maps were computed from spatially normalised images warped directly from the native space. This means a mapping between intensities of native Prisma and Trio scans was

learned instead of native Prisma and coregistered Trio scans. Examination of regional mean FA confirmed this hypothesis. The neural network was able to best project Prisma scans to coregistered Trio scans. However, coregistered Trio scans introduced a bias that lowered the FA values. This might be caused by intensity interpolation and rotation of b-vectors during the coregistration. The mapping learned by the neural network also adapted to that bias. Consequently, FA values after CNN harmonisation were much lower than the target Trio FA intensity domain. For clinical applications, researchers will have to take this and the model complexity into consideration to choose an appropriate harmonisation method.

For the control subject of both databases, MUSHAC and CENTER-TBI, the harmonisation methods are qualitatively comparable. While all methods could effectively minimise the RMSE between scans, the CNN with multiple pathways was overall more successful in reducing the scanner differences. This was particularly evident for RISH features of higher order (RISH_2 and RISH_4). For b_0 and RISH_0 images the benefits of CNN Multi-Path harmonisation was not as prominent in the MUSHAC database than in the CENTER-TBI database. As described above, data for one outlier subject were not harmonised well, which shows the sensitivity of neural networks to such outliers.

5.4.2 Enhancement of Scaling Maps Through Post-Processing

The standard approach of linear RISH feature scaling involves computing scaling maps for the expected values of RISH features (i.e. the average image per scanner) in template space and apply these to SH coefficients after backprojection to native space. Inherently, there is a disconnection between RISH features maps, that are the sum of squared SH coefficients, and the SH images themselves. While RISH_0 feature maps are only derived from one SH image, higher order RISH features condense information from several SH images. With less structured signal distribution in higher order SH images, the RISH features might be less representative of SH coefficients on a voxel-wise level. Denoising images with a median filter generally can remove speckles. For RISH feature scaling maps this means removing voxels that are disconnected between RISH and SH images and avoid arbitrary scaling. This was observed to be beneficial on RISH_4 images, however, merely to recover the original inter-scanner differences. A similar observation was made for Gaussian filter, but its blurring effect was disadvantageous for more structured images such as the b_0 image. Although the benefits of denoising scaling maps were marginal, this experiment revealed a general flaw in the concept of SH coefficient scaling via RISH features.

The linear RISH feature scaling was later on improved by including an adaptive approach to compute the scaling maps [187]. This idea of selecting subjects to derive the scaling

factors was applied to CENTER-TBI data. Furthermore, this was extended by computing weighted averages of feature maps, instead of a strict subselection of a few subjects, to derive scaling maps. Overall, these approaches resulted in very similar performances. However, the selection of a few subjects did not bring the previously reported boost in performance. This may be explained by the previous observation that an optimal performance was achieved with a minimum of 16 training subjects used to calculate expected RISH feature values [126]. Including more subjects leads to smoother average images hence less speckled scaling maps, which further would improve scaling of higher order RISH features. Indeed, selection of the three most similar subjects resulted in noisier scaling maps and was a disadvantage for RISH_4 images. This is coherent with the previous observation that smoother RISH_4 scaling maps led to better data harmonisation. The voxel-wise weighting to compute scaling factors was most beneficial for RISH_4 feature maps, as it is most capable to adapt to the unstructured intensities for each individual test subject.

5.4.3 Application to Traumatic Brain Injury Data

Both differences between scanners and patient groups with good and poor outcome were too subtle to be measured on a region-wise level. This made it difficult to estimate the impact of mapping intensities via ROI-wise scaling. The more complex neural network resulted in the adversary effect of reduced FA. As mentioned above this is likely due to a learned mapping between Prisma and coregistered Trio scans, rather than the raw Trio data. Important is also the observation that the FA difference between patients with good and poor outcome slightly increased. The models presented in the literature and in this chapter are usually trained on control subjects. This could particularly be problematic for TBI, as the model learns to project diffusion metrics that appear healthy, but possibly will fail for outlier intensities associated with subtle pathology. This could lead to an artificial enhancement of differences between patient groups.

The examined methods are based on a one-to-one mapping between scanners. This is not necessarily feasible for a large multi-centre study, such as CENTER-TBI, that includes more than ten acquisition sites. Not only is it computationally expensive to compute one model for each mapping, but there is not yet a consensus how to pick the optimal reference site to harmonise all data to. Regarding the CENTER-TBI database, there are many other challenges to overcome. By design only a maximum of nine healthy controls were scanned at each site. With at least 16 subjects required for an optimal mapping between scanners [126], the Linear-RISH harmonisation might not be suitable, in particular, as Cambridge was the only CENTER-TBI site that collected data for the same controls on all available scanners. Both the acquisition protocol and the scanned volunteers will differ between imaging sites.

This and the limited number of control subjects makes it much harder - or even impossible - to match control scans across sites, which hampers the computation of adequate scaling maps.

5.4.4 Future Work

The experiments presented in this chapter have shown that neural networks have the capability to learn a mapping between different scanner domains. However, the models presented here and compared previously [238] rely on matched scans of travelling volunteers, which often are not available for multi-centre studies. The Linear-RISH harmonisation has been successfully applied to larger databases without travelling volunteers [32, 225], but require at least 16 well matched control subjects [126]. There have been efforts to compute mappings between RISH feature images via deep learning models that do not rely on directly matched control subjects [127], however, these are based on Cycle-GANs [285], which are difficult to train. In addition, any machine learning model will perform best on its training data. To avoid a bias, training subjects would need to be excluded from any following analysis. For a study with limited number of healthy volunteers, such as CENTER-TBI, control subjects could be used for data harmonisation, but not for any subsequent analysis. Future investigation will need to look into harmonisation methods focusing on limited data and explore ideas how to incorporate non-imaging patient data to avoid introducing any biases.

A further development is the simultaneous harmonisation of scans from multiple sites. The VAE introduced by Moyer et al. [182] learns scanner independent representations between several scanners, hence could be useful for a multi-centre study. The auto-encoder outperformed Linear-RISH harmonisation, however, could still be improved by using a larger receptive field (patches instead on directly adjacent neighbour voxels). Furthermore, it could benefit from including multiple-pathways for the different SH orders. The VAE could also be extended to use more variables than just the site/scanner identifier as covariates. This could possibly allow to include patient data as well. A recent study has successfully shown the joint reconstruction of diffusion metrics from multiple-sites, while simultaneously reducing scanner biases [244]. Such a model could possibly be adapted to fit control and TBI patient data together to minimise inter-site variation while preserving information from pathological diffusion.

Besides the use of more data, covariates or different neural network architectures, the cost function to train the model is also worth investigating more in the future. The FSCNet optimised the network parameters not only by computing the loss function between SH images but also RISH feature maps, that were computed from harmonised signal throughout the training phase. Likewise, the VAE designed by Moyer et al. [182] incorporated both

differences between SH and DWI images.

5.5 Chapter Summary

This chapter examined different methods for diffusion MRI harmonisation. The results showed that neural networks are best at learning a non-linear mapping between scanners based on SH representation. Thereby, it was highlighted that different orders of SH benefit from being processed separately on multiple neural network pathways. The experiments also revealed potential pitfalls, such as, for example, relying the harmonisation on coregistered data. Here, it was shown that coregistering data from the reference site changed the diffusion signal. Mapping data from the source site to such modified data from the reference site may result in a skewed data harmonisation. Further investigation is needed to make harmonisation a reliable processing step for clinical multi-centre databases.

Chapter 6

Lesions Analysis in Severe TBI

6.1 Introduction

Examination of brain lesions is a broad research field for many different pathologies. In contrast to other brain diseases that may show distinct patterns across patients (e.g. multiple sclerosis or stroke), TBI lesions are very heterogeneous showing a wide range of severity [21]. Dependent on the strength and site of the external force impacting the head, traumatically induced pathologies may include focal lesion masses, such as contusions and haematomas, or diffuse axonal and microvascular injury. Although injuries are predominately focal or diffuse, most lesions consist of both components [172] and vary greatly in size, shape and location. To add to the complexity, secondary brain injuries evolve long time after the traumatic incident [40] leading to changing lesion patterns and eventually to late-life consequences [25, 172, 257]. Understanding lesion formation and evolution throughout different stages post-injury is crucial to gain insight into traumatic head injuries and ultimately improve patient outcome. Therefore, many studies have been undertaken to identify and characterise TBI lesions [73, 170, 271]. Detecting abnormalities is indeed a very challenging task in itself and the first step to analyse lesions. Therefore, much justifiable research has been conducted to detect brain lesions on MR images [13, 110, 159]. However, far less attention has been given to subsequent analysis strategies to derive clinically valuable information from segmented brain lesions. Once the lesions have been manually or automatically outlined on the brain scans, the total lesion volumes can be quickly computed. Gaining a comprehensive understanding of the pathology, however, requires further effort. Imagine two lesions on different scans with the exact same *total* volume. While one lesion may affect one large region within the brain, the other could be separated in smaller clusters more widely distributed. Many smaller lesion clusters may have a different effect than one large lesion, despite in sum having the

same volume. Nonetheless, in a simple volumetric assessment both lesions would appear to be equal. Besides that, a lesion in one brain region could be more detrimental than a lesion of similar size found in another. Thus, lesions may be better characterised by assessing their volume as well as their location and distribution within the brain.

6.1.1 Motivation from a Clinical Research Perspective

The following section describes some examples of TBI research in which lesion location was linked to clinical variables, such as patient outcome. The focus lies here on visible lesions (e.g. contusions or oedema apparent on FLAIR scans), that could be annotated and visually located by the naked eye. Latent lesions that manifest for example as changed diffusion metrics based on ROI analysis or voxel-wise statistics will not be considered.

Moen et al. [180] evaluated the prognostic value of visible traumatic axonal injuries by identifying and counting lesions on several MRI contrasts. Analysing 64[1] severe TBI patients, revealed that the number of DWI lesions and the volume of FLAIR lesions in the CC, brainstem and thalamus were predictive of patient outcome measured by the GOSE. The number of cortical contusions on MRI scans were more informative for moderate TBI patients [180]. Another study that included longitudinal data with acute and chronic scans from 16 TBI patients found lesion volumes correlated with patient outcome. It was discovered that only larger TBI lesions in the temporal lobes, as found on GRE and FLAIR scans, were connected to brain tissue volume loss. Such lesions could be associated with worse neuropsychological outcome.[2] In contrast, tissue atrophy was not tied to frontally located lesions and their volumes seemed to have no impact on patients outcome scores. This study, however, was only based on 16 TBI patients, with 11 lesions detected in the frontal lobes [169]. A comparative study used different MRI contrasts to qualitatively assess lesion locations in 38 TBI patients. For this, the brains were subdivided into a *superficial*, *deep* and *posterior fossa* brain zone. Lesions were only associated with a single zone, and in case the lesion crossed more than one zone it was assigned to the zone containing most of the lesion. This revealed that volumes, number and regional distribution of early T2w and FLAIR lesions were useful to distinguish patients with good and bad outcome (dichotomised by *Glasgow outcome scale* [GOS][3] score). The latter was associated with larger volumes and higher lesion numbers. Particularly, median total volumes of lesions in the superficial zone were most consistently disseminating between different patient outcomes. The authors hy-

[1] the study included a total of 128 moderate or severe TBI cases
[2] oral and written *Symbol Digit Modalities Test*
[3] 1: Death, 2: Neurodegenerative state, 3: Severe disability, 4: Moderate disability, 5: Good recovery

pothesised that this might be because the superficial zone comprises the four major brain lobes, hence includes most of the brain volumes. However, there was not differentiation between subcortical WM and cortical areas. Susceptibility weighted images were found to be beneficial to highlight intraparenchymal injury, but weakly linked to outcome scores [35]. Furthermore, neurocognitive effects[4] of TBI were studied in 71 patients. Lesion location was estimated qualitatively by associating them to different brain zones, such as for example prefrontal cortex or temporal lobe. Large prefrontal lesions were linked to an impaired performance in the neurocognitive tests assessed. Patients with lesions that were located in and beyond the frontal regions had the worst test performance [155].

While head trauma is an incidental event, the brain injury is often progressing during the first hours after impact. This is due to expansion or development of new hemorrhagic lesions after cerebral contusions. Since this could result in tissue with likely unrecoverable loss of function, it is important to study lesion development to gain better understanding of secondary brain injury [143]. Rehman et al. [213] have studied the progression of hemorrhagic contusions in 246 patients with TBI that underwent an initial and follow-up (within 24h) *computer tomography* (CT) scan. Contusion volumes on CT scans were manually estimated by measuring lesion diameters and counting the affected image slices (this approach to estimate lesion volume is also known as the ABC method). Progression was defined as an increase of $> 30\%$ of the initial volume [213]. Furthermore, the location, laterality and multiplicity of the lesion was assessed through visual inspection. Hemorrhagic contusion progression could be associated to initially large contusion volumes (> 20 ml) and low score on the GCS. A retrospective study of 491 TBI patients with admission and follow-up CT scans within 72 hours revealed a progression of intra-parenchymal haemorrhage for three quarters of the patients. Thereby, the lesion expanded on average by approximately 62% (~ 5 ml). Several factors were found to contribute to the rate of haemorrhage progression, but lesion volume on the admission CT seemed most predictive [29].

6.1.2 General Concept for Lesion Localisation

Figure 6.1 displays an example of a MR scans showing a TBI with two different lesion clusters that were annotated manually (Figure 6.1 B). The lesion is clearly visible on different MRI contrast images such as FLAIR (Figure 6.1 A) and T1w (Figure 6.1 C). In order to locate the lesion, it has to be seen in context with anatomical regions within the brain. Once the brain is divided into meaningful areas, the overlap between the lesion annotation and each region provides information about the lesion's location. One way to define this overlap is to

[4] performance in gambling test

calculate the relative volume of a lesion that lies within a particular brain region (e.g. 20% of the lesion is within left middle frontal gyrus). To automatically parcellate the brain in subregions, the multi-atlas tool MALP-EM was chosen as it has been shown to deal well with distorted brain anatomy [151]. Nonetheless, lesion pathology often introduces an additional challenge for parcellation algorithms. Brain parcellation via MALP-EM is based on locally non-linear registration of multiple atlases and a subsequent refinement of the regions by means of expectation maximisation. While this latter adjustment is highly advantageous to improve region segmentation, it is also sensitive to abnormal intensity profiles in the images. Therefore, in the presence of lesions the algorithm could be hindered in its ability to parcellate the brain into anatomical plausible regions leading to corrupted segmentations (see right top corner within the brain Figure 6.1 D). Comparing any lesion segmentation to the erroneous region atlas would be flawed and lead to incorrect results. Alternative brain parcellation methods, such as FreeSurfer's recon-all, would also suffer from distorted parcellations since these algorithms were not designed to cope with large lesions. Thus, instead of using an accurate but lesion-sensitive algorithm, the idea is to use an estimate for the brain parcellation that was derived by projecting the MALP-EM atlas of a healthy subject population onto a patient's brain. While not perfectly accurate, this parcellation will be more robust. This idea of atlas registration to a lesioned brain is not new and has been employed to localise stroke lesions [137], however, it has not been widely adopted for TBI research, in which lesions are often located via visual inspection. Further methodological details will be described later on.

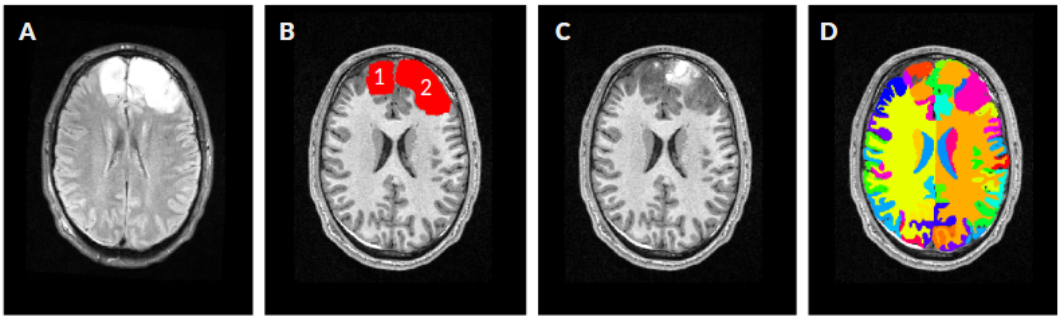


Figure 6.1: Problematic of Automated Brain Parcellation via MALP-EM in Presence of Lesions. On a FLAIR scan (A) of lesioned brain, two clusters were identified and manually annotated (B). Since the lesion is also prevalent on the T1w scan (C), the brain region parcellation, driven by the T1w image intensities, are disturbed (D different regions highlighted in different colours, see right top corner within the brain).

6.1.3 Assessment of Lesion Progression

Lesions observed in patients suffering from TBI are vastly heterogeneous. To gain a better understanding of pathological changes over time, lesions need to be compared across longitudinal scans. The usual approach for analysis is the use of manual lesion annotations to evaluate the lesion burden at different time points. Previous studies mostly assessed lesions by measuring the total lesion volume [213, 277] and counting the number of lesions as seen on different imaging contrast (FLAIR, SWI, etc.) [11, 78]. However, comparing total lesion volumes over time may oversimplify the evolution of individual lesion clusters. For example the tissue recovery from a small lesion could go unnoticed when a different lesion cluster on the same scans substantially grew. Measuring the total lesion volume for comparison across time points would show an increase in lesion burden, while failing to reflect the salvageable tissue. Therefore, it may be important to examine individual lesion clusters independently to capture more nuanced changes. For a direct comparison, single lesion clusters will need to be matched between scans. In other words, for each lesion cluster found on the initial scan the corresponding part in the follow-up scan needs to be identified. With increasing numbers of patients in recent imaging studies, the tracking of each lesion part across longitudinal scans can be a tedious task, and calls for automated approaches to measure lesion changes. Automated lesion matching across scans, however, bears many challenges. At first, a spatial correspondence between both scans needs to be established. Two scans of a healthy subject could be registered rigidly, assuming the scans were acquired temporally close, as no brain tissue atrophy is expected. However, TBI patients may display different pathological patterns at different stages, as brain tissue swelling and deformation can take place at different time points post-injury [108]. Thus, a deformable registration is often more successful to bring brain tissues in alignment. Secondly, lesions in the two scans will need to be matched depending on their spatial overlap. This is simple for one large, corresponding lesion in both scans, and is fairly straight-forward to generalise for several lesion clusters as long as their number remains constant across scans. However, this becomes quickly more complicated when lesions dissolve or newly form, or one big lesion partially recovers such that is appears to split into several smaller lesion clusters at a later stage. Algorithms to match lesions will need to account for this variability in lesion evolution.

6.1.4 Aims

The aim of this chapter is to introduce two approaches to localise TBI lesions and estimate longitudinal lesion changes. Both methods were designed to be fully automated, without the need for an expert to decide via visual inspection where the lesion are located or how

they evolve. Applied to a cohort of severe TBI patients, the localisation algorithm was used to examine in which brain areas TBI lesions are found predominantly. Experiments will also investigate differences for cohorts scanned on two different scanners as a case study for potential variability in multi-centre data. Furthermore, locations of lesions will be assessed for three different time points after injury. Lesion characteristics, such as volume and location, will be estimated both in template and subject-space to compare the validity of both approaches. A registration-based algorithm for matching lesions between longitudinal scans will be introduced. This methods will then be validated qualitatively and used to examine lesion changes over time in a subset of TBI patients.

6.2 Data & Methods

6.2.1 Severe TBI Database

Following experiments made use of a severe TBI patient cohort scanned on a Siemens Prisma and Verio scanner. For both scanners T1w image were acquired with the same sequence parameters as follows: $TR = 2300\,ms$, $TE = 2.98\,ms$, $TI = 900\,ms$ and FOV read = 256 mm with percent phase FOV= 93.8% and a flip angle of 9 degrees. The FLAIR scans were acquired with $TR = 27840\,ms$, $TE = 95\,ms$, $TI = 2500\,ms$ and FOV read = 224 mm with percent phase FOV= 80.8% and a flip angle of 9 degrees. Various lesion types and artefacts were manually annotated by clinicians. However, the focus laid on the contusion core and oedema, seen on FLAIR images, as these have been delineated within this database most frequently. The database included 112 patients with scans at different time points after injury. Patient were scanned fairly evenly on both scanners (Trio: 59 patients, Verio: 53 patients). In total there were 187 scan sessions with both T1w and FLAIR images. All of these had a manually annotated oedema lesion. Contusion cores were annotated on 183 of those FLAIR scans. Manual segmentations were generate by a clinical expert outlining both lesion types (contusion and oedema) on FLAIR scans by means of *ImSeg*, an in-house tool[5] to view and annotate medical images. The selected data were processed via the structural MRI pipeline (Section 2.3), so that the bias field corrected and spatially normalised T1w images as well as coregistered FLAIR sequences were available afterwards. Further processing steps for lesion localisation and longitudinal matching will be described below.

[5] originally developed by Microsoft Research

6.2.2 Localisation of Lesions

The inverse of the transformation from T1w space to the Cam-CAN template from the structural MRI pipeline was used to backproject the template's MALP-EM ROI atlas to each individual T1w scan. Thereby, nearest neighbour interpolation was chosen to keep the integer ROI labels intact. The transformation for coregistering FLAIR to T1w images was applied to the manually created binary lesion maps of FLAIR contusion and oedema, again using nearest neighbour interpolation. Both projections led to the alignment of an uncorrupted ROI atlas (i.e. the healthy subjects Cam-CAN ROI atlas) and the binary lesion maps within a subject's T1w image space. In practice the backprojection of the ROIs could lead to part of the lesion not being covered by the atlas. To prevent this, the ROIs in the template atlas were fully expanded to cover the whole FOV within template space. For this, each voxel without ROI label was assigned the one of the closest ROI (estimated via Euclidean distance). Backprojecting this expanded template atlas ensured full coverage of any lesion. Figure 6.2 shows two examples of the original MALP-EM atlas along side the backprojected Cam-CAN ROI atlas.

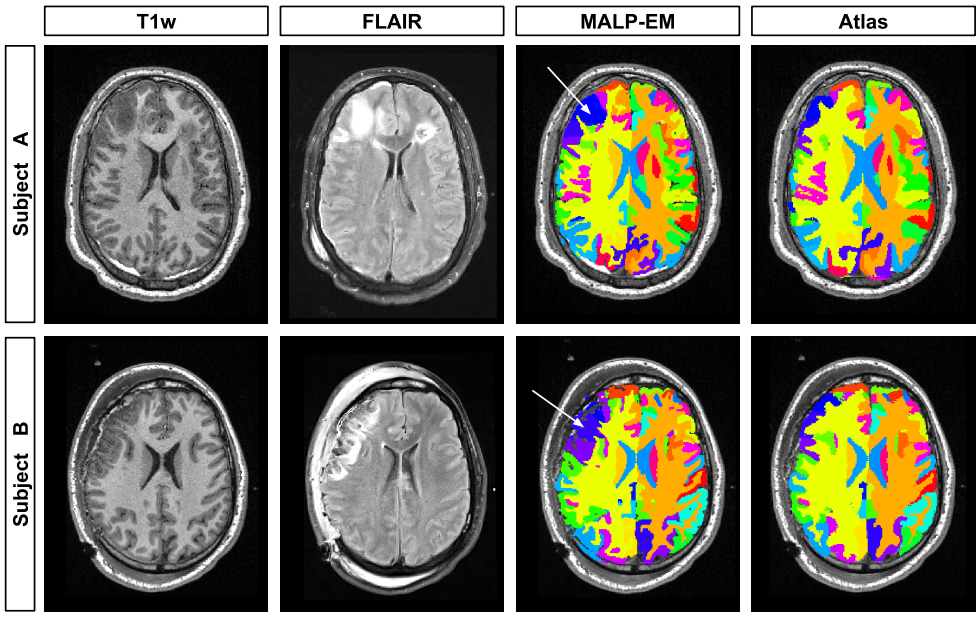


Figure 6.2: Comparison of Original and Projected Parcellation. Each row shows axial slices of two example subjects. The first column show the T1w MR scan that was used for brain parcellation. The corresponding FLAIR images (second column) clearly shows the contusion core and oedema (e.g. hyper-intense regions in the frontal brain regions). When subdividing the lesioned brains into brain regions via MALP-EM (third column) an erroneous parcellation could be observed (white arrows). In contrast, backprojecting the parcellation of the Cam-CAN template shows a less lesion-affected parcellation, albeit not as precise (fourth column)

An in-house python implementation took the aligned atlas and lesion maps as input and first split the binary lesion annotation into individual lesion clusters by means of connected component analysis (measure.label algorithm from the skimage python library). Then, for each lesion cluster the *overlap* with all 138 MALP-EM ROIs was computed as the ratio of the lesion volume within the particular ROI and the total lesion volume. Besides this lesion overlap, further characteristics were computed automatically. This included the *total lesion volume*, the *average lesion cluster volume*, the *number of lesion clusters* as well as the *lesion distribution*, computed as the average distance between the centres of lesion clusters to one and another.

For an alternative group-wise comparison, lesions were also directly projected from native FLAIR space to Cam-CAN space. This was achieved by applying transformations from coregistration and spatial normalisation simultaneously to avoid multiple interpolation steps (one nearest neighbour interpolation). To quantify the occurrence of lesions in template space, the projected binary lesion maps were averaged to compute a probability map, indicating each location's relative frequency to have been labelled as FLAIR contusion core or oedema. The mean value of these probability maps within the ROIs of the Cam-CAN MALP-EM atlas were used to determine a region's *group-wise lesion burden/occurrence*. A comparison of the lesion burden at different time points after the incidental brain injury was used to estimate a lesion type's *progression* within the cohort. All projections were performed with the antsApplyTransforms tool.

6.2.3 Longitudinal Lesion Matching

Subjects from the severe TBI database were selected when they have had a FLAIR contusion core and oedema manually annotated on both a hyper-acute (scan within 72 hours) and an acute follow-up scan approximately one week post injury. These inclusion criteria were fulfilled by 21 patients (DPI [mean±std]: 1.3±0.6 or 7.6±2.1 for the hyper-acute and acute stage, respectively).

Figure 6.3 schematically visualises the registration-based approach for longitudinal lesion matching. At first, FLAIR scans were rigidly registered to the corresponding T1w image at each scan session (Figure 6.3: 1. Coregistration to T1w Images). This alignment has been completed as part of the structural MRI processing pipeline as discussed earlier (Section 2.3) Afterwards the T1w scans from the hyper-acute and acute phases were registered to one and another to create a subject-specific template (Figure 6.3: 2. T1w Image Registration). For this, the follow-up scan was at first rigidly aligned to the early scan. Afterwards both scans were repeatedly registered to the common template, that was the average of the aligned images of the previous registration stage. The iterative registration

process entailed three affine and three deformable registrations. Within the subset of 21 patients, initial and follow-up scans showed strong anatomical deviance, which is why a deformable registration was required to ensure an optimal spatial overlap between lesions in both scans. Eventually, the NCC was computed between the last registered images and the particular subject-specific template, to flag any cases of failed registration. High NCC scores (average NCC=0.953±0.028) and a visual quality check confirmed the success of the template creation process. The final deformable transformations were then applied together with the rigid transformation, found during FLAIR image coregistration, to the FLAIR lesions maps. This projected the delineated lesion labels with a single interpolation step from native FLAIR to subject-specific template space. Subsequently, projected lesions were split into individual *lesion clusters*, to categorise all unconnected parts of the binary lesion map. Thereafter, the overlap between each lesion cluster within both scans was computed to match individual lesion parts to one another (Figure 6.3: 3. Lesion Projection & Matching). All matched scan pairs were visually inspected and no obvious errors, such as wrongly assigned labels between corresponding lesion cluster, were observed. Nine lesion clusters in total were excluded, as their volume was 1 mm^3 (which corresponded to one voxel in this database) because the registration based matching algorithm was not expected to have a single voxel precision. Eventually, the volume of each separate lesion cluster in native T1w scan space could be associated with the matched clusters (Figure 6.3: 4. Derive Information from Matched Lesions). This allowed to measure the volume change, unbiased by the deformable registration, for individual clusters between scans. Lesion volumes were derived from binary lesions in T1w space as lesion maps are normalised to isotropic voxel space.

6.3 Results

6.3.1 Group-Wise Lesion Burden Across Scanners

After spatial normalisation of T1w images and the projection of FLAIR lesions to Cam-CAN template space the binary maps were averaged for both scanners separately to detect predispositions in both patient groups (in the following referred to as *Trio* and *Verio* patients/scans depending on the MR scanner model). The results presented include 89 and 96 lesion maps from Trio scans for contusion core and oedema, respectively. From Verio scans, 92 contusion core and 89 contusion oedema annotations were found. The probability maps of the projected lesion maps are displayed in Figure 6.4. While patient groups on both scanners showed contusion cores predominantly in frontal brain regions, subjects scanned on Trio had a higher lesion burden in the right hemisphere (i.e. left hand side on the image

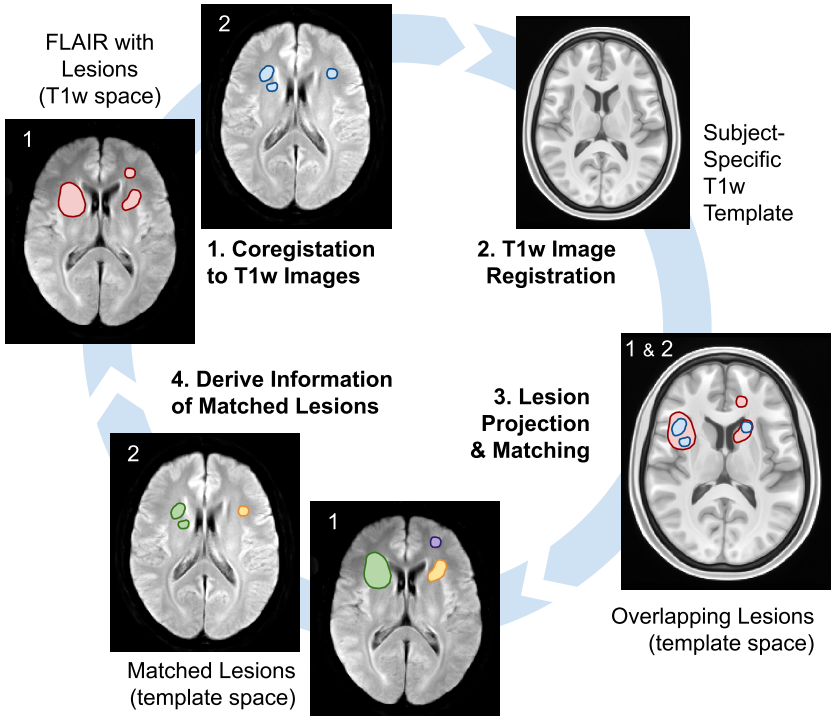


Figure 6.3: Schematic Overview of Lesion Matching Between Longitudinal Scans. At first, the sequence images with annotate lesions (here FLAIR) were linearly coregistered to corresponding T1w images. Secondly, a subject-specific template was generated by iteratively registering T1w images from two different time points. This was followed by projecting lesion annotations (red & blue outlines) directly from native to template space. Spatial alignment allowed to computing overlaps between all different lesion clusters from longitudinal scans to match lesions (yellow & green) and identify single lesions (purple). Eventually, this association between scans was propagated to lesion annotations on T1w images to derive characteristics for matched lesion clusters.

with radiological orientation). In contrast, Verio patients revealed a higher lesion density in the left hemisphere (right side of image). Contusion oedema were more evenly distributed in both hemispheres for Trio and Verio patients, but also much wider spread than contusion cores. Additionally, a much higher occurrence of oedema was found in the Verio patients.

For the two patient groups from both scanners, the highest occurrence of contusion cores was detected in the anterior orbital gyrus (frontal lobe). Thereby, lesions were found slightly more often on the right (ROI #45: 6.4%) then on the left (ROI #46 4.0%) for Trio patients, and vice versa for Verio scans (right ROI #45: 6.7%, left ROI #46: 8.4%), Generally, a tendency for higher lesion density within the right frontal lobe was observed for Trio scans with comparable high mean probability value in the medial orbital (ROI #81: 2.8%), inferior frontal (ROI #97: 2.8%) and posterior orbital (ROI #111: 3.0%) gyri. In contrast, contusion cores on Verio scans were more likely to be in the left frontal lobe, in partic-

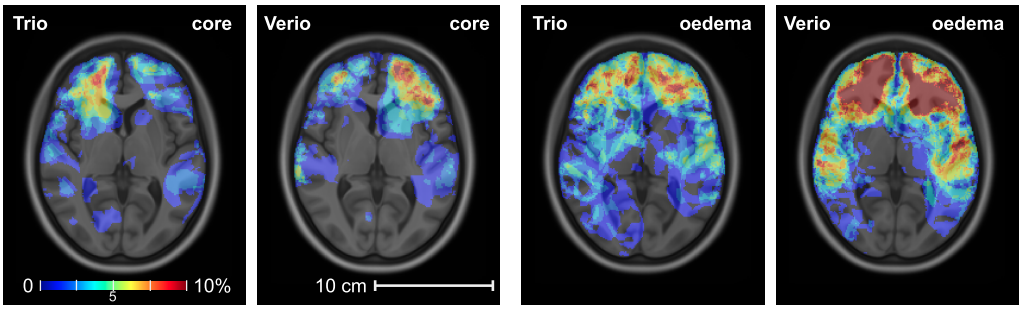


Figure 6.4: Distribution of FLAIR Contusion Core and Oedema. All axial slices show the Cam-CAN T1w template overlayed with the lesion probability maps displayed as jet map ranging between 0-10%. The two images on the left hand side show the distribution of the contusion cores for all subjects and time points separated by scanner (Trio & Verio). The Trio patients had a higher lesion occurrence in the right frontal lobe (left image side) compared to Verio patients, that displayed a stronger contusion core presence in the left frontal lobe (right image side). For both patient groups lesions are dominant in the frontal areas, but also spread across temporal lobes. The two images on the right show the contusion oedema probability maps for both patient groups. Overall, oedema is much more widely distributed over the whole brain and fairly evenly present in both hemispheres than core lesions. Noticeable is the high occurrence of oedema in the frontal lobes for patients imaged on the Verio scanner. (radiological orientation)

ular in the lateral orbital (ROI #72: 4.9%) and inferior frontal (ROI #98: 4.0%) gyri. These findings are in coherence with the previous visual observations in Figure 6.4. Besides this opposite laterality, contusion cores were also found in the medial-frontal cortex (ROI #76: Trio=5.6%, Verio=10.7%) and the posterior orbital gyrus (ROI #112: Trio=5.3%, Verio=11.1%) of the frontal lobe in either of the two patient groups. Furthermore, the left (planum polare - ROI #114: Trio=6.5%, Verio=11.9%) and right (temporal pole - ROI #134: Trio=5.5%, Verio=11.9%) temporal lobes were affected.

Oedema was found to be mostly present in the right frontal lobe (anterior orbital gyrus - ROI #45: 5.3%) for Trio patients and more often in the left inferior frontal gyrus (frontal lobe - ROI #98: 12.7%) for Verio patients. Overall, contusion oedema had occurred much more in Verio than Trio patients (lesion burden for top five ROIs: Trio$\sim$5%, Verio$\sim$10%).

6.3.2 Lesion Volume Progression After TBI

All scans were subdivided into three classes according to the acquisition time point measured. These were the *hyper-acute* ($\leq$72 hours post-injury), *acute* ($\sim$1-2 weeks post-injury) and *subacute* ($\sim$1-3 months post-injury) stages. The selected number of scans was fairly comparable, ranging from 44 to 71 (Table 6.1), for all three defined time windows. The number of scans were mostly balanced between both scanners (Trio:40, Verio: 60%) across all time points. The comparison of total lesion volumes within Cam-CAN space revealed

that contusion cores were much smaller (≈10-15 cm^3) than oedemas (≈30-55 cm^3) during all three observed stages. Both, core and oedema seemed to be on average largest during the acute phase. However, while the contusion core volume increased by approximately 15% from the hyper-acute to the acute stage, oedema expanded on average by 81% within the same time frame (Table 6.1).

Table 6.1: Overview of Lesion Volumes for Different Time Windows after TBI

	Scan Time Window	Avg. DPI mean ± std	DPI Range [min, max]	# Scans (Trio/Verio)	Volume* [cm^3] mean ± std
Core	hyper-acute	1.8 ± 0.8	[0, 3]	59 (35/24)	12.8 ± 19.9
	acute	7.1 ± 2.5	[4, 12]	71 (31/40)	14.7 ± 22.2
	sub-acute	24.6 ± 12.8	[13, 64]	44 (19/25)	10.4 ± 16.7
Oedema	hyper-acute	1.8 ± 0.8	[0, 3]	61 (37/24)	30.6 ± 39.6
	acute	7.0 ± 2.4	[4, 12]	71 (32/39)	55.4 ± 54.7
	sub-acute	26.2± 16.4	[13, 91]	43 (19/24)	30.9 ± 42.9

*Total volume in Cam-CAN template space

Figure 6.5 shows the occurrence of both lesion core and oedema throughout the different time points on one axial slice. Both lesion types were most prevalent in the acute phase, where they spread over frontal and temporal lobes. Although already present in the hyper-acute stage, in particular oedemas seemed to be much more dominant during the acute phase, indicating a delayed development. The strong presence of oedemas appeared to diminish with more time passing after the incident of the traumatic injury. These visual observations were in agreement with the previous quantitative analysis.

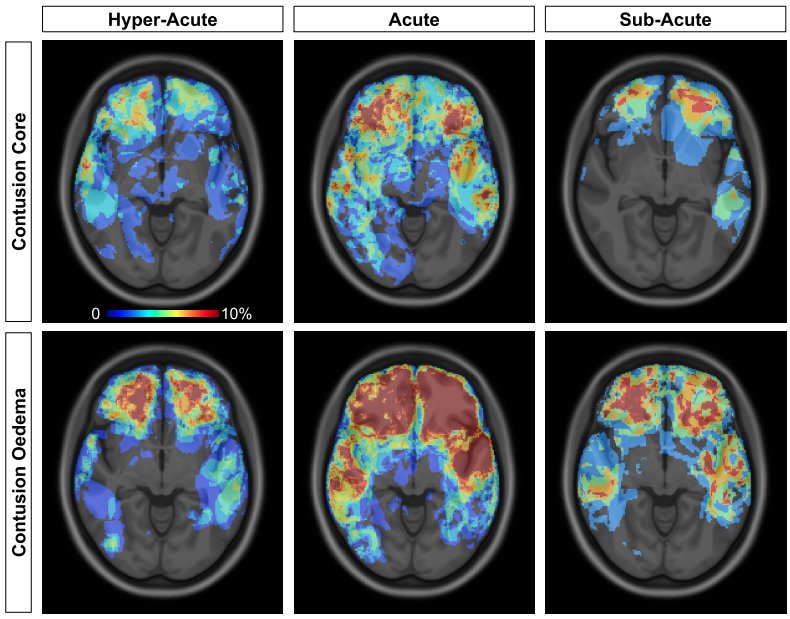


Figure 6.5: Progression of Contusion Core and Oedema after TBI. The probability maps of contusion core and oedema at different time points are displayed as overlays of the Cam-CAN T1w template. Both lesion types were most prominent during the acute phase. Lesions occurred predominantly in the frontal and temporal lobes. All jet colour maps ranging between 0-10%. (radiological orientation)

A subset of 21 subjects (11/10 Trio/Verio; 7/14 female/male; average age: 42.6 ± 17.2; initial GCS: 6.4 ± 3.1) was selected which had an available manual annotation of contusion core and oedema at the hyper-acute (mean ± std [min, max] DPI = 1.3 ± 0.6 [1, 3]) and acute (DPI = 7.6 ± 2.0 [5, 11]) stages. For this subset of subjects the total core lesion volume (measured in T1w space) was on average indistinguishable (paired t-test: p=0.5990) between the hyper-acute (19.5 ± 25.5 cm^3) and the acute (18.5 ± 24.9 cm^3) phases. In contrast the total lesion volume of oedema showed a substantial (paired t-test: p<0.001) growth between the hyper-acute (40.1 ± 47.5) and acute (58.4 ± 55.6) stages. Thirteen subjects had on average a 37.0% (range: [0.2%-98.5%]) smaller core lesion volume in the acute phase than in comparison in the hyper-acute phase. Six other subjects showed a moderate 22.9% (range: [2.1%-76.2%]) core lesion growth from the hyper-acute to the acute

phase. Only two patients experienced substantial core volume increase (465.7% and 587.4%). Five subjects showed a small shrinkage of oedema volume (17.4% on average, range: [0.2%-40.2%]), while the other 17 patients experienced an oedema growth (105.9% on average, range: [0.3%-310.4%]) from the hyper-acute to the acute phase. Not only did more patients have an increased oedema lesion, but also the volume expansion was greater (76.5%) than that for core tissue.[6] Measured by the GOS scores, there was a tendency of better outcome for patients with contusion core lesion shrinkage (number of patients: 13, average GOS: 3.6) in comparison to patients with core lesion volume growth (number of patients: 8, average GOS: 2.8). However, low sample numbers and small differences did not manifest statistically (t-test: p-value=0.1067). A similar trend was not observed for oedema volume change, remarking again the unbalanced ratio between patients with oedema volume loss (5) and gain (17).

6.3.3 Subject-Wise Cross-Scanner Comparison of Lesions

Location. After locating all contusion cores and oedemas for all subjects separately (Section 6.2.2), the relative frequency of lesion occurrence was estimated within each region. For this, a patient's brain region was counted as affected, whenever any part of the lesion was found within that parcellated area. This low threshold (effectively one voxel would be enough to count the lesion as affected) was chosen as any other arbitrary threshold (e.g. 50% of the lesion must be in a region to be affected) would need to be justified empirically. It is acknowledged here that this makes the count of affected lesions very sensitive. For both Trio and Verio scans the most frequently affected regions were the left (ROI #16: Trio = 36.8% scans, Verio = 37.1% scans) and right cerebral WM (ROI #17: Trio=32.9% scans, Verio=28.1% scans). This means, more than a third of the patients' scans had at least a part of the contusion core within WM regions. Approximately, 7-8% of the scans displayed a contusion core within cortical regions such as the orbital, frontal and temporal gyri. On average, approximately 30%-60% of the contusion core volume was laying in one of those mentioned regions. This was cohesive with the group-wise analysis showing a higher lesion load in frontal and temporal areas of the brains. Likewise, oedemas most frequently were found in the cerebral WM for cohorts scanned on both scanners (left ROI #16: Trio = 41.4% scans, Verio = 35.4% scans; and right ROI #17: Trio = 31.0% scans, Verio = 33.8% scans). Furthermore, the right superior frontal (ROI #121: Trio = 10.2% scans, Verio = 9.4% scans) and the left middle frontal (ROI #78: Trio & Verio = 8.8% scans) gyri were affected by oedema.

[6] excluding the two patients with major core volume increase

Volumes & Distribution. The average volumes of contusion core lesion clusters were similar for both scanners (Trio: 1.5±2.6 cm^3, Verio: 1.5±2.8 cm^3; t-test: $p = 0.4006$). In contrast to that, oedemas for patients scanned on Verio (6.5±6.3 cm^3) were larger than for the cohort imaged on Trio (3.9±6.3 cm^3; t-test: $p = 0.0009$). Oedemas were generally larger than lesion cores. The distance between individual lesion clusters on single scans averaged between approximately 5.5-5.9 cm and were comparable for both scanners (Trio & Verio) and lesion types (core & oedema). Unsurprisingly, the mean FLAIR intensities within the annotated lesions were different for the two scanners, as no intensity harmonisation was performed.[7] Generally, oedema displayed higher intensities than contusion cores. However, mean intensities were lower on Trio than on Verio for both core lesions (Trio=310.5, Verio=399.4) and oedema (Trio=411.6, Verio=500.0). Interestingly, mean intensities in Verio cores seemed closer to Trio oedema ($\|\Delta\| = 12.2$)[8] than both the differences between mean intensities of core ($\|\Delta\| = 88.9$) and oedema ($\|\Delta\| = 88.4$) across scanners.

6.3.4 Subject-Wise Lesion Characteristics on Longitudinal Scans

Location. Counting the instances where at least one part of the MALP-EM ROI was affected by core lesion tissue (overlap between atlas ROI and annotated lesions), showed again that the most affected regions where left and right cerebral WM throughout all three imaging stages (hyper-acute/acute/sub-acute).

Approximately 30-40% of the scans showed an overlap with cerebral WM in at least one brain hemisphere. For both regions the average lesion overlap was highest during the acute phase (ROI #16: 61.0%, ROI #17: 56.6%, see Figure 6.6 left), and lower for hyper-acute ROI #16: 53.5%, ROI #17: 45.9%) and sub-acute stages (ROI #16: 52.1%, ROI #17: 47.7%). In both brain hemispheres the superior frontal gyrus (left: ROI #121, right: ROI #122) as well as the temporal pole (left: ROI #133, right: #134) were affected with similar frequency in 7-11% of the scans. Both superior frontal gyrus regions displayed a higher average overlap in the hyper-acute phase (ROI #121: 49.6%, ROI #122: 45.3%) than in the acute (ROI #121: 32.2%, ROI #122: 35.5%) or sub-acute phases (ROI #121: 39.0%, ROI #122: 18.8%). Although lesions overlapping with the left temporal pole (ROI #133) were progressively larger (hyper-acute: 53.5%, acute: 61.8%, sub-acute: 67.8%) no obvious difference was found for the right counterpart (ROI #134 ∼55-58%, see Figure 6.6 left).

Likewise 30-40% of the scans showed oedema overlapping with the cerebral WM regions. Generally, proportional overlap of oedema lesions with particular regions tended to be lower

[7] Adjusting intensities across scanners in lesioned brains in non-trivial and standard approaches such as histogram matching can have an adversary effect.

[8] $\|\Delta\|$ representing here the absolute intensity difference.

than that found for contusion cores. This was a side effect of oedema being larger and more widespread across several regions. Nonetheless, the average overlap with cerebral WM seemed slightly larger in the acute (ROI #16: 53.5%, ROI #17: 53.3%) than in the hyper-acute stage (ROI #16: 48.9%, ROI #17: 49.0%, see Figure 6.6 right), which could indicate a growth in lesion volume after 72 hours post-injury. Other areas affected by oedema were the left posterior orbital gyrus (ROI #112) and the superior frontal gyri (ROI #121 & #122). For these regions, however, the overlap tended to decrease with more time after injury (e.g. ROI #112 - hyper-acute: 21.0%, acute: 5.8%, sub-acute: 9.1% or ROI #122 - hyper-acute: 35.2%, acute: 29.5%, sub-acute: 24.3%). A similar but less prominent tendency was observed for the left middle frontal gyrus (ROI #78 - hyper-acute: 31.7%, acute: 31.2%, sub-acute: 27.6%).

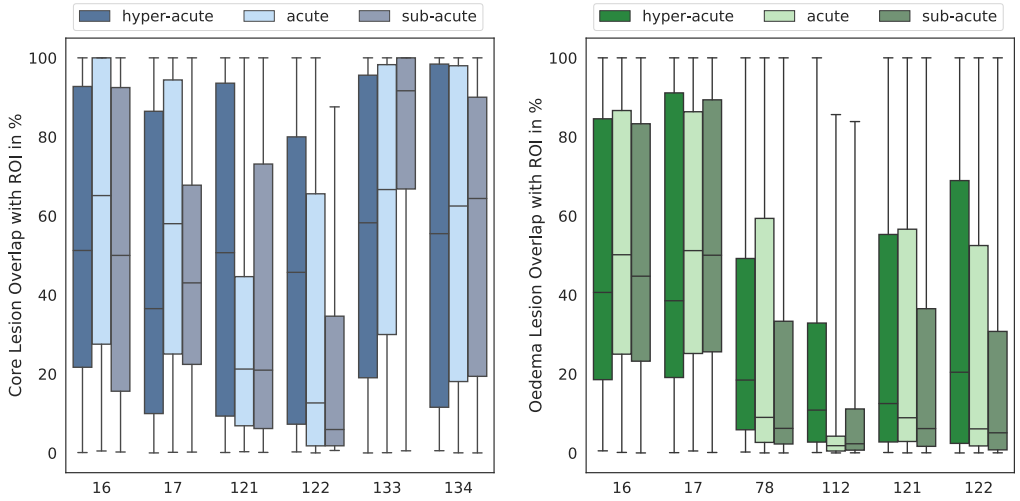


Figure 6.6: Lesion Overlap at Different Phases Post-Injury. **Left**: Average overlap of core lesions with MALP-EM ROIs. Displayed are lesions that were most frequently affected. The overlaps vary depending on the anatomical regions. While core lesion seem to overlap with cerebral WM (16 & 17) more during the acute phase, superior frontal gyri (121 & 122) show a decrease in overlap from hyper-acute to acute stage. An opposite effect was observed for temporal poles (133 & 134), for which the overlap appeared to steadily increase with time after injury. **Right**: On average oedema overlap was highest for WM regions (16 & 17), which seemed larger after the hyper-acute phase. Other areas such as the middle frontal gyrus (78), the posterior orbital (112) gyrus and the superior frontal (121 & 122) gyri were affect less by oedema. For these the overlap seemed to diminish with more time after injury. Whiskers show the full range of ROI overlap. X-axes show the different MALP-EM ROIs.

Volumes. For all three time windows lesion volumes strongly varied for different subjects (Figure 6.7). Table 6.2 summarises the characteristics of the lesions found in T1w space for all subjects. The total volume of the contusion core lesions during the hyper-acute phase averaged to 12.9 cm^3. Considering the whole cohort together, the core lesion volume seemed to slightly increase during the acute time window (14.3 cm^3) and then shrunk below the

initial state with more time elapsed after the injury (sub-acute: 10.0 cm^3). A statistical significance of total volume difference for the three time points, however, was not detected (ANOVA: p=0.5044).

Comparing the individual core lesion clusters showed similar average volumes during the hyper-acute (1.2 cm^3) and acute (1.3 cm^3) imaging time points, however, the maximum number of clusters increased. This could indicate that a tendency for larger total core lesion volumes may be attributed to development of new small core lesions, rather than growth of the initial lesions. In contrast, individual clusters seemed to have become bigger during the sub-acute phase (1.8 cm^3), but both the average and maximum number of clusters was decreased (6 and 19, respectively). This could suggest that some core lesion tissue was salvaged, while other core lesions expanded. The observed trend for the change in cluster volumes was not statistically supported (ANOVA: p=0.4037).

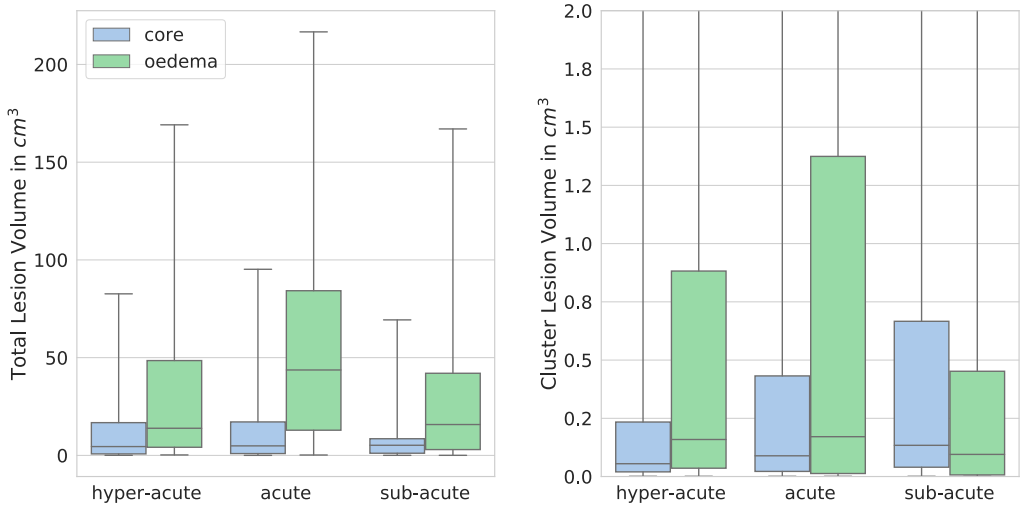


Figure 6.7: Longitudinal Comparison of Contusion Lesion Volumes Post-Injury. **Left**: Total volumes of manually annotated FLAIR contusion cores seemed to be mostly stable across all three considered time points centring around 10-15 cm^3 and not exceeding 100 cm^3. In contrast oedema volume seemed to expand ($\sim$50 cm^3) after 72 hours during the acute phase, but residing to the initial state in the sub-acute phase ($\sim$30 cm^3). Oedema volumes were found to be on average much lager than core volumes. **Right**: Volumes of individual core clusters appeared to grow marginally with more time elapsed post injury. Comparable to total volumes, single clusters of oedema expanded during the acute phase, but resolved at the later sub-acute stage. Note, y-axis for cluster volumes was restricted to 2 cm^3 for visualisation purposes (90% of all found clusters had a volume below 2 cm^3). Whiskers show the full range of lesion volumes.

A growth of total contusion oedema volume from hyper-acute (30.5 cm^3) to acute (54.9 cm^3) stage was much more prominent than for core tissue (ANOVA: p=0.0022, post-hoc t-test: p=0.0025). The scans acquired during the sub-acute stage showed on average lesion volume (30.5 cm^3) that was closer to the initial state when imaged within 72 hours (ANOVA:

Table 6.2: Lesion Characteristics After TBI

	Scan Time Window	Total Volume[1] mean[median]±std	Cluster Volume[2] mean[median]±std	# Clusters[3] mean[median, max]
Core	hyper-acute	12.9[4.5]±19.3	1.2[0.06]±5.7	11[9, 34]
	acute	14.3[4.9]±20.2	1.3[0.09]±5.3	11[9, 39]
	sub-acute	10.0[5.2]±14.9	1.8[0.13]±6.3	6[4, 19]
Oedema	hyper-acute	30.5[13.9]±38.5	3.6[0.16]±14.5	8[6, 24]
	acute	54.9[43.7]±50.9	6.0[0.17]±20.4	9[8, 31]
	sub-acute	30.5[15.8]±39.7	4.3[0.10]±16.9	7[6, 29]

[1]Total lesion volume [cm^3] found in T1w space; [2]Average volume [cm^3] of lesion clusters found in T1w space; [3]Minimum number of found cluster for all stages was 1

p=0.0022, post-hoc t-test p=0.9909). The observation (Table 6.2) of an early growth of oedema in the acute phase and its subsidence in the sub-acute stage (ANOVA: p=0.0022, post-hoc t-test: acute vs. sub-acute p=0.0010) is in alignment with the previous results from the group-wise analysis in Cam-CAN space. Both the maximum number of clusters and their average lesion volume increased during the acute phase, and followed the trend of the total lesion volumes. However, a statistically significant volume change of individual lesion clusters across the three time points was not found (ANOVA: p=0.0629).

A linear correlation between total lesion volume and DPI was neither found for core lesions (Spearman correlation coefficient ρ=-0.04, p=0.61) nor for oedema (Spearman correlation coefficient ρ=-0.02, p=0.74). This underlines the findings in Figure 6.7 and Table 6.2, where a lesion growth was observed between hyper-acute and acute phase, but lesion volumes seemed to shrink between acute and sub-acute stage. However, a correlation between the volume of lesion clusters against DPI could not be rejected (Spearman correlation coefficient: Core: ρ=0.12, p<0.001; Oedema: ρ=-0.07, p=0.01). The results from volume differences for different time points and the correlation with DPI emphasised the non-linear development of lesions. These did not monotonically grow or subside within the first three months post injury.

6.3.5 Automated Lesion Matching

The same 21 subjects as before (Section 6.3.2) were used to match individual lesion clusters between the hyper-acute and acute scans (Section 6.2.3). Three selected subjects with lesion annotations in subject-specific template space are displayed in Figure 6.8. This shows not only the successful alignment between subjects despite severe pathology, but also the

association of individual lesion clusters between the two imaging time points. Despite the change in size or vanishing contusion, large contusion clusters were easily matched across scans (Patient B & C). For smaller lesions, the matching might be challenging when clusters do not show an overlap. On the other hand, the algorithm was able to separate contusion core clusters in spite of close proximity (Patient A - blue & green). The association of oedema clusters between hyper-acute and acute scans was consistently successful, even with many new forming clusters (Patient A - green, blue & red).

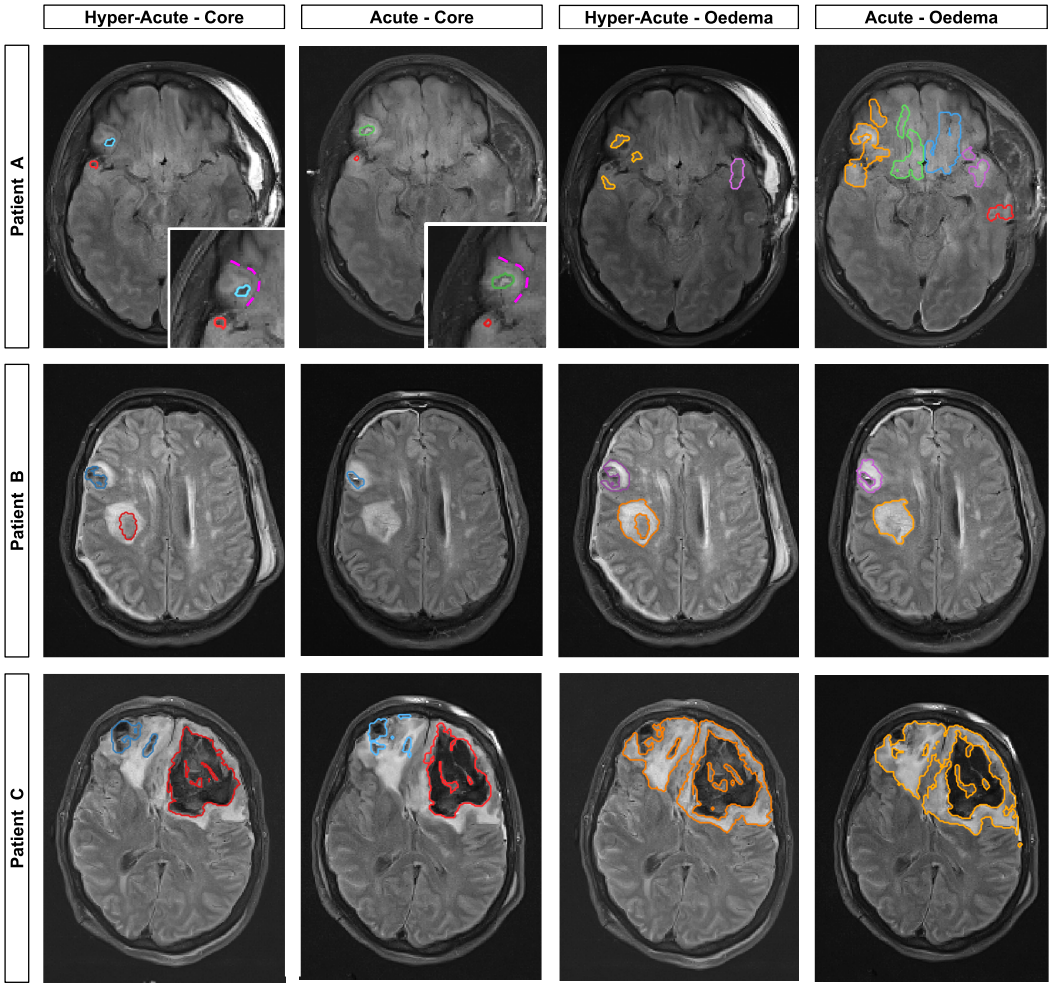


Figure 6.8: Examples of Matched Contusions between Hyper-Acute and Acute Scans. Each row shows a different patient's FLAIR scans from the two different time points with outlines of the manual lesion annotations for contusion core and oedemas. Outlines with the same colours indicate matched lesion clusters for each individual lesion type (e.g. hyper-acute and acute core clusters matched). **Patient A**: One contusion cluster was successfully matched (red). Two other clusters were recognised as independent (blue & green) since there was no spatial overlap. The detailed scope view highlights the spatial separation between both clusters, nonetheless, both clusters might belong to the same pathological structure. New oedema formed after the hyper-acute scan (red, green & blue), while the existing clusters (orange & purple) were matched to the right clusters at a later stage. Note, the orange cluster appears only separated on this axial slice. **Patient B**: Despite vanishing core lesion (red), the 2nd cluster was matched adequate (blue). Successful matching of oedema clusters. **Patient C**: Both two core lesions (red & blue) and growing oedema were associated correctly across scans. All scans in subject-specific template space, image intensities were minimally adjusted for better visualisation. (radiological orientation)

Table 6.3: Volume Changes of Matched TBI Contusion Clusters. Volumes displayed mean [min, max]

	Type	# Clusters$^{\times}$	Hyper-Acute Volume	Acute Volume	Growth^{+}
Core	All	326	1.6cm^3	2.3cm^3	58.0%
	Forming	65 ∼20.0%	-	0.1cm^3 [3.0mm^3, 3.1cm^3]	-
	Growing	60 ∼18.5%	2.5cm^3 [13.0mm^3, 62.0cm^3]	3.2cm^3 [21.0 mm^3, 68.3cm^3]	128.8%
	Shrinking	47 ∼14.5%	5.3cm^3 [19.0mm^3, 68.8cm^3]	4.0cm^3 [4.0 mm^3, 57.7cm^3]	-32.4%
	Vanishing	154 ∼47.0%	0.1cm^3 [2.0mm^3, 1.7cm^3]	-	-
Oedema	All	253	5.2cm^3	7.3cm^3	96.7%
	Forming	91 ∼36.0%	-	0.4cm^3 [2.0mm^3, 5.4cm^3]	-
	Growing	59 ∼23.5%	12.5cm^3 [39.0mm^3, 162.6cm^3]	19.0cm^3 [94mm^3, 210.8cm^3]	137.0%
	Shrinking	18 ∼7.0%	4.8cm^3 [72mm^3, 4.3cm^3]	3.9cm^3 [16.0mm^3, 4.3cm^3]	-35.6%
	Vanishing	85 ∼33.5%	0.2cm^3 [2.0mm^3, 1.5cm^3]	-	-

All: Considering all lesion cluster volumes of 21 severe TBI patients. **Forming**: Lesion clusters that were *not* present during hyper-acute phase but were visible on the acute scans. **Growing**: Matched lesion clusters with *larger* volumes during acute than hyper-acute phase. **Shrinking**: Matched lesion clusters with *smaller* volumes during acute than hyper-acute phase (negative growth). **Vanishing**: Lesion clusters in the hyper-acute phase, that were not observed in the acute phase. $^{\times}$Number of clusters found on either of the hyper-acute and acute scan. $^{+}$Growth was defined as the volume difference between matched clusters divided by the hyper-acute cluster volume, negative growth indicated a loss in lesion volume.

Table 6.3 summarises the volume and the growth of the matched lesion clusters. Almost half (∼47.0%) of the contusion core clusters vanished from the hyper-acute to the acute phase. Noteworthy is that those were overall small lesions with average volumes of 100 mm^3 (and not exceeding 1.7 cm^3). Approximately a fifth (∼20.0%) of all core clusters were not present during the hyper-acute phase, but developed later, hence, were only visible on the acute MR scans. Similarly to the vanishing clusters, the forming cluster volume averaged to 100 mm^3, however, with a slightly higher maximum volume (3.1 cm^3). The rest of the core clusters were either growing or shrinking, whereas slightly more clusters seemed to grow (∼18.5%) than shrink (∼14.5%). While clusters shrunk about a third (∼32.4%) in volume, growth was much more pronounced (∼128.8%). On average, growing clusters were initially smaller (2.5 cm^3) than shrinking clusters (5.3 cm^3) during the hyper-acute stage. This difference was almost eradicated at the acute imaging phase (3.2 cm^3 and 4.0 cm^3 for growing and shrinking cluster volumes, respectively). Although core lesion clusters grew more than they shrunk, much more clusters vanished than newly formed between the hyper-acute and acute phases. So, although there is a measurable change for individual lesion clusters the overall burden due to contusion lesions might be consistent over time.
Fewer clusters were found for oedemas, which is likely due to their larger volumes and less fragmented distribution. Indeed, during both imaging stages, oedema lesions were much larger than core cluster volumes. With a substantial growth between hyper-acute and acute injury phases. Roughly a third (∼36.0%) of the oedema clusters formed after 72 hours post injury and were only observed on the acute scan. Another third (∼33.5%) were present on the hyper-acute scans, but were not observed on the acute scan anymore. Both of these cluster types showed low average volumes, whereas newly forming clusters (mean: 400 mm^3, max: 5.4 cm^3) tend to be slightly larger than vanishing clusters (mean: 200 mm^3, max: 1.5 cm^3). Oedema lesions were observed to grow more (137.0%) than they shrank (35.0%). With comparable numbers and volumes of forming and vanishing clusters, oedema lesion burden seemed to increase after 72 hours.

Table 6.4 shows the total volume changes of contusion cores and oedemas between the hyper-acute and acute phase for the individual subjects. The majority of subjects (16) showed a gain in oedema volume, whereas seven patients experienced an increase in total core lesion tissue (Patients 1-7). The other nine patients (Patients 9-17) showed lower core lesion volumes. Shrinking oedema lesions were observed only in five patients (Patients 8 & 18-21), which was accompanied by contusion core growth exclusively for one single patient (Patient 8). This highlighted the tendency of growing oedema and shrinking core lesions volumes between the hyper-acute and the acute stage after injury.

Table 6.4: Patient-Wise Changes of Contusion Core and Oedema Volumes. Total volume differences (Δ) between hyper-acute and acute stage listed in cm^3 except for marked ($^{\times}$) values, that are shown in mm^3.

	Core				**Oedema**			
Patient	Δ	Growth	Loss	Total^{+}	Δ	Growth	Loss	Total^{+}
1	2.2	5.6%	2.4%	3.3%	26.7	21.7%	0.3%	21.4%
2	4.7	15.2%	2.9%	12.3%	15.5	31.7%	8.3%	23.4%
3	12.6	17.1%	1.9%	15.2%	47.6	29.6%	1.5%	28.1%
4	0.3	13.4%	11.3%	2.1%	14.2	44.2%	0.4%	43.8%
5	9.0	470.3%	4.7%	465.6%	12.3	46.0%	0.0%	46.0%
6	4.2	619.9%	32.4%	587.4%	**4.1**	**149.0%**	**67.0%**	**81.9%**
7	**0.1**	**144.6%**	**68.4%**	**76.2%**	32.0	220.8%	0.0%	220.7%
8	6.0	29.7%	1.6%	28.1%	-1.1	5.1%	7.2%	-2.1%
9	-2$^{\times}$	3.5%	7.0%	-3.5%	3$^{\times}$	41.6%	41.4%	0.3%
10	-1.8	3.2%	31.7%	-28.4%	1.6	9.2%	4.1%	5.1%
11	-6.9	2.7%	29.2%	-26.5%	41.9	64.0%	0.6%	63.4%
12	-2.9	1.4%	18.7%	-17.2%	32.0	97.2%	0.0%	97.2%
13	**-0.2**	**24.9%**	**46.8%**	**-21.9%**	8.5	148.5%	22.6%	125.9%
14	-5.2	0.5%	70.2%	-69.7%	19.0	147.5%	6.9%	140.6%
15	-3.5	4.2%	24.3%	-20.0%	36.4	227.8%	7.2%	220.6%
16	-34.9	1.0%	46.8%	-45.8%	59.5	270.2%	5.2%	265.0%
17	-3.9	9.3%	33.8%	-24.5%	39.4	314.1%	3.8%	310.4%
18	-1.4	1.5%	100.0%	-98.5%	-4.0	0.0%	40.2%	-40.2%
19	-44$^{\times}$	0.0%	66.7%	-66.7%	**-0.1**	**27.2%**	**61.3%**	**-34.1%**
20	-0.5	2.7%	61.1%	-58.4%	-0.3	29.8%	40.0%	-10.2%
21	-31$^{\times}$	2.1%	2.3%	-0.2%	-0.2	0.9%	1.0%	-0.2%

Sorting according to total lesion change (*Total*): **Patients 1-7**: Growth of both core and oedema lesions. **Subject 8**: Contusion core growth but oedema shrinkage. **Patients 9-17**: Shrinking contusion core, but growing oedema. **Patients 18-21**: Both lesions decreased overall in size. $^{+}$Positive and negative values represent a total gain or loss of volume, respectively.

For many patients, the total lesion volume change reflected the underlying changes of lesion clusters. This was particularly true for patients where either lesion volume gain or loss, was dominant compared to the other. Exemplary was the oedema lesion development of patient 4 (dominant oedema growth): The growth of the oedema volume (44.2%) was much more prominent as its observed volume loss (0.4%). Therefore, the total oedema volume change (43.8%) accurately represents the overall development of oedema size. Similarly for patient 14 (dominant core loss), for which the contusion core clusters hardly grew (0.5%), but the core lesion volume substantially decreased (70.2%). Hence, the total measured volume loss (-69.7%) accurately reflected the overall change in contusion core size. However, some patients' overall lesion volume failed to represent both the gain and loss of lesion tissue (printed bold in Table 6.4). Total contusion core change for patient 7 was observed as 76.2%. Although there was a great growth in lesion volume (144.6%), the substantial decrease in contusion core lesions of 68.4% was missed when measuring only the total lesion change (0.1 cm^3). Similarly, patient 6 experienced a substantial growth in oedema lesion volume (149.0%). Although this was accompanied by a considerable loss of oedema lesion volume (67.0%) as well, this was not reflected in the total oedema lesion change of 4.1 cm^3 (81.9%). In contrast, patient 13 showed a total loss of contusion core volume of -21.9%, while missing a lesion growth of 24.9%. Likewise, a total decrease of oedema volume was observed for patient 19 (-34.1%), which reflected a combination of substantial oedema lesion growth (27.2%) and loss (61.3%). Noteworthy, apart from patient 6 the total core/oedema lesion volume differences of the four mentioned examples are below 0.3 cm^3.

6.4 Discussion

6.4.1 Summary of Findings

The analysis of group-wise lesion burden showed opposite uni-lateral dominance of contusion cores in the two TBI patient cohorts scanned either on Trio or Verio. Oedema was found to be more evenly distributed in both hemispheres for both patient groups, however, more strongly present in patients scanned on Verio. While categorising the patients according to the used MR scanner was arbitrary and their opposing lesion locations is likely due to coincidental difference in trauma impact site, it highlights the heterogeneity of TBI patient cohorts. This advocates for an analysis of large patient database across multiple imaging sites to capture the heterogeneity in TBI cohorts and avoid any site-specific biases. For example, a machine learning model for automated lesion segmentation trained on data from one scanner could learn that lesions are predominantly in one hemisphere. But when applied

to a different cohort with lesions in both or the opposite hemisphere (as seen in core lesions for Verio and Trio scans, Figure 6.4), the segmentation algorithm may perform poorly due to its site-specific bias introduced by the skewed training data. Contusion cores were mostly found in the frontal and temporal lobes. This is in agreement with previous reports of contusions appearing in brain tissue that comes in contact with irregular bony protuberances of the skull [172].

Lesion volumes were found to be largest during the acute phase (1-2 weeks post injury). Examining a subset of 21 subjects showed that the total lesion volume of contusion cores either grew moderately or often also decreased in a later stage post-injury. In contrast, most of the selected subject experienced a growth of oedema volume, which strongly expanded between the hyper-acute and acute stage. This is in agreement with previous studies, which also have reported a delayed appearance of oedema after a few hours post injury and the increase to its peak volume a few days later [108, 203]. There are two types of oedema, cytotoxic and vasogenic. The former is characterised by an increase in water content in the intra-cellular space caused by dysfunctional ion pumps that fail to regulate cell osmolarity. An unbalanced ion gradient results in influx of extracellular water into the cells [56]. In contrast, vasogenic oedema is caused by water movement from the vasculature system to extracellular compartment due to an impaired *blood–brain barrier* (BBB). Both types of oedema occur in TBI in a biphasic profile. Vasogenic oedema emerges within the first few hours after TBI. This is followed by cytotoxic oedema which develops more slowly over few days and can persisted for up to 2 weeks [56]. The increased oedema volume during the acute phase for the available data (time frame after injury: mean [min, max] = 7 [4, 12] DPI, see Table 6.1), may indicate the observed oedema is of cytotoxic nature. Examination of the BBB has shown, that its permeability is highest at 4-6 hours after TBI and the commences to close over the week following TBI. The recovery of the integrity of the BBB is the reason why oedema stop increasing after a few days and subside eventually during the sub-acute phase [56].

The subject-wise analysis of lesion location showed that cerebral WM regions were most frequently affected. One reason for this is the fact that the WM regions defined by the MALP-EM atlas are spanning the largest area within the brain. Therefore, any fairly large lesion will almost certainly be at least partially within one of the cerebral WM regions. The specificity of WM lesions could be increased by incorporating a more detailed subdivision of the WM areas. Besides WM, contusion cores were observed in frontal, temporal and orbital gyri. Hee Kwak et al. [97] have recently reported that patients with frontal lesions

showed higher agitation (e.g. aggressiveness, restlessness or mood swings measured by on the Agitated Behaviour Scale) and poorer performance in executive and emotional functions (measured by Wisconsin Card Sorting Test, for more detail the reader is referred to [83, 97]). These metrics were not available for the cohort presented in this chapter. Future investigations will have to tie together radiological observations with behavioural data. Furthermore, contusions in the temporal lobes may be associated with with worse functional outcome after six months [275].

While the average lesion cluster volume of contusion cores was similar for both patient cohorts, oedema volumes were larger for the patient group scanned on Verio. The differences in volumes might be explained by the fact that oedema volumes were largest during the acute phase, and proportionally more acute scans were acquired on Verio (∼45%) than on Trio (∼36%). Since MRI intensities are not standardised, another difference was found between FLAIR intensity distributions between both lesion types and scanners. Such potential bias between scanners will need to be considered and prevented for predictive models built for multi-centre image analysis. For example, lesion segmentation tools dependent on image intensities may fail to segment lesion adequately and lesion volumes might be under- or overestimated.

Comparing the lesions for the three time points, showed that overlap with cerebral WM was largest during the acute phase for both lesion types examined. This could indicate a growth in lesion volume. However, the overlap with smaller ROIs (e.g. middle frontal gyrus) simultaneously decreased. Potentially, the lesion could shrink within smaller regions, increasing the relative portion of lesion within the large lesion *without* actually increasing the overlap. Therefore, the metric of lesion overlap with certain regions needs to be considered together with lesion volumes.

The lesion matching algorithm introduced in this chapter showed potential, however, it might face challenges in the presence of small lesion clusters. For almost half of the annotated core lesion clusters on hyper-acute scans no corresponding cluster on the follow-up scan was found. Although this might indicate actually dissolving contusion core lesions, it also highlights the difficulty to match lesions of small volumes. Smaller lesions are more challenging to associate across scans, since their size lowers the chance of a spatial overlap after spatial alignment of both scans. Furthermore, small lesions are much harder to detect and may have been missed at the annotation stage. Small lesions, however, are less likely to progress [1], hence may not lead to major behavioural deficits. Nonetheless, the automated lesion matching between hyper-acute and acute scans was successful for larger lesions. Since

contusion core lesions are naturally smaller than oedema, the former are more prone to inaccurate matching. Therefore, analysis based on the suggested automated lesion matching approach should take into account lesion type and size. Matching individual lesion clusters across longitudinal scans could provide better insight of lesion volume gain and loss. This separation, rather than a single value for total volume change, may potentially be more suitable to differentiate between patients with good and poor outcome.

6.4.2 Limitations of Study

Localisation. The results presented showed that TBI lesions could be located automatically within the brain, however, this remains a challenging problem. The current approach is strongly dependent on the successful spatial normalisation of a patient's brain scan to the Cam-CAN template. For severely deformed brains this could be error prone, and results will need to be considered cautiously. Assuming the spatial normalisation has worked sufficiently well, the backprojected template ROI atlas shows an average brain anatomy and does not perfectly align with subjects individual structures. Since it is unlikely that a patient underwent a MRI scan prior, but timely close to the TBI, the comparison of a lesioned brain to its true healthy anatomy is practically impossible. The suggested approach builds on the assumption that lesions affect tissue without any occurring deformations of anatomical regions. For example, a lesion in one region could cause swelling, pushing other tissue compartments aside, but leaves them otherwise unaffected. Such dislocated, but still functional brain tissue would not be captured by projecting a healthy atlas onto the lesion scans, but would require complex physical modelling of brain tissue properties. Nonetheless, for a general estimate of lesion location, particularly for larger lesions, superimposing a healthy brain atlas has worked sufficiently for the available data.

Quantification. With the successful alignment of lesion maps and anatomical atlas, quantifying lesion localisation is far from trivial. One of the metrics introduced in this chapter is the lesion overlap with an anatomical region, which was defined as the volumetric proportion of a lesion cluster associated with a particular ROI. However, this measure does not represent the lesion burden by itself. For example two lesions could both half overlap (50% proportion within ROI) with the same region, but the region's lesion burden could be drastically different depending on the lesion size. The straight-forward way to measure a region's lesion burden would be to compute the percentage of a region that is covered by a lesion, but because lesions clusters often have low volumes in comparison to the anatomical region, the measured lesion burden can be vanishing small. Thus, subtle differentiation of small lesions would be numerically difficult. Quantifying lesion burden will need to be a

combination of lesion location, size, shape and co-occurrence with other lesion types.

Lesion Matching. The approach suggested for lesion matching was based on image registration. This is challenging for heavily distorted brains, and will have a direct impact on the lesion matching performance. While mostly unproblematic for large lesions, small lesion clusters are more prone to not be paired across corresponding scans. As shown earlier, small contusion cores that were spatially apart were classified as independent core clusters, while they might actually belong to the same pathology (Figure 6.8 *Patient A - Core*). The insufficient overlap of lesion clusters could be a result of inadequate spatial correspondence or interpolation errors of lesion masks introduced by the registration. Findings suggested that vanishing and newly forming lesions have all lesion sizes below 0.5 cm^3. While it makes sense, that rather smaller than larger lesion were salvageable and formed at a later stage, this indicated strong built-in biases against matching small lesions successfully. Firstly, small lesions are generally harder to detect and to annotate, which is why small lesions might be missing on the initial or follow-up scan. If indeed detected on both scans, finding an overlap of small lesions via registering is much less robust than for large volumes. So the suggested approach of lesion matching via spatial correspondence, might only be robust for larger lesions. Although tracking the progression of individual lesion clusters was more informative than measuring the change in total lesion volume for a few subjects, the results did not clearly support the need for lesion matching. Considering the results as proof of concept, future experiments will need to focus on larger cohorts to understand the benefits of matching individual lesion clusters for overall lesion progression.

Sample Size. The presented analysis was based on manually lesion annotations. Although many scans were visually inspected to label lesions, finding patients that fulfilled all requirements (e.g. delineated contusion core and oedema on both hyper-acute and acute scans) quickly reduced the sample size (here: 21). While this allowed to test algorithmic concepts and per case studies, it is challenging to draw more than simple and general conclusions. This holds especially true for TBI cohorts as patients are very different to one and another. High variability of lesion characteristics (size and location), demographics (age at injury, sex) as well as clinical observations (initial injury severity, responsiveness) hinder correlation of lesion findings to patient outcome.

6.4.3 Future Work

As mentioned, the heterogeneity of TBI patients makes it hard to find similar patterns of lesion development within a small cohort. Since manual lesion labelling is time-consuming, the generation of an annotated dataset is limited. However, with an automated lesion seg-

mentation tool [125] it will be possible create lesion maps for a larger TBI database. The same concepts for lesion localisation and matching will then be applicable to analyse lesion changes in an extended patient cohort. This will also help to further find strength and weaknesses of the algorithms here suggested to examine lesions. Besides expanding the available data, both the lesion localisation and the lesion matching algorithms can further be improved. One obvious choice for future enhancement will be to use a more detailed WM atlas, that allows a more refined localisation of sub-cortical lesions. In addition, a more advanced algorithm could be developed that builds upon MALP-EM, that allows full brain parcellation in the presence of lesions. One option would be to build a predictive model, that learns to *remove* lesions [65, 205] on a scan such that MALP-EM could be applied to pseudo-healthy brains without corrupting the underlying multi-atlas projection by the present lesions. However, removal of pathology via inpainting is a challenging open research question and this approach would still rely on MALP-EM's time consuming parcellation. Alternatively, a model could be build that predicts ROI parcellations [153, 158] and lesion segmentation simultaneously [57, 85].

The lesion matching algorithm associated clusters when there was any overlap, however, as discussed previously small lesions may not overlap despite belonging to the same pathology. One possibility to avoid missed matching could be to connect smaller lesion clusters when they are in very close proximity and have not been associated with any other clusters. A threshold for maximum distance would need to be found empirically. Another enhancement could be to build a graph connecting lesion clusters within the same and across longitudinal scans to find a better correspondence between clusters. Although it is unclear how the different number of lesions would affect the structure of such a graph. Matching lesion clusters by proximity and graph metrics would need extensive visual validation. The lesion matching algorithm could be tested on synthetic data, however, generating artificial TBI lesions that is challenging due to heterogeneous location and appearance. Furthermore, TBI lesions may cause brain tissue deformation, why is difficult to simulate. So, synthetic lesions possibly do not reflect the complexity found in real TBI data.

Eventually, information from both lesion location and changes over time could be combined to fully characterise lesions. This could further be used to predict outcome measures in a larger cohort to distinguish patients with different outcomes. One limitation of the experiments presented in the chapter is the link to outcome metrics (i.e. GOS), which might be to crude. Future investigations should also focus on examining correlations between lesion derived metrics and more nuanced cognitive behaviour or symptom scale. For example the Rivermead Post Concussion Symptoms Questionnaire [62, 131] could be used. Comprised of 16 questions concerned with physical, cognitive and behavioural symptoms it provides a

more detailed insight in a patient's well-being. A larger sample size in combination with a more elaborated analysis of cognitive deficits and behavioural changes will also be more conclusive, whether longitudinal matching of individual lesion clusters actually will have a clinical impact.

6.5 Chapter Summary

In this chapter, two concepts for a fully automated lesion analysis were examined: Firstly, atlas registration-based lesion localisation (which has previously applied to other disorders, such as stroke [137]), and secondly, lesion cluster matching between longitudinal MR scans. Their potential was explored for manually annotated contusion cores and oedema found on FLAIR scans of TBI patients. Lesions were located based on their overlap with regions of the MALP-EM atlas. Visual inspection has shown the superiority of the backprojection of a healthy atlas over a using the MALP-EM parcellation calculated for individual patients. Although less detailed, the former showed better robustness in the presence of lesions visible on T1w scans. Contusion cores and oedema were found to be predominantly in frontal and temporal regions. Lesion volumes were generally larger for oedema, and seemed to expand after the hyper-acute phase, but may grow or resolved later on. The overlap of lesions with specific atlas ROIs showed a strong bias towards region size (e.g. large cerebral WM regions were most frequently affected) and more precise metrics may need to be derived to gain a better insight. The lesion matching between hyper-acute to acute MR scans showed promising results for larger lesions. However, small lesions were found to be more prone to failed matching. Despite the difficulty of accurate segmentation, lesions of lower volumes naturally have smaller chance of overlap between with corresponding lesions, which may limit the application of these methods. Enhancing the association of lesions between scans through more sophisticated methods could help to improve the matching of smaller lesions.

Chapter 7

Summary

With the recognition of the potential of large sample sizes and the previously unseen availability of data, the field of neuroimaging is experiencing a shift towards big data science. To answer more complex clinical research questions, new collaborations are formed to built multi-centre studies that acquire thousands of MRI scans. This goes hand in hand with new innovative analysis tools, such as deep learning, that both demand and foster the collection of large datasets. While they need many samples to fit adequately to a data distribution, they also allow to learn complex relations in data that could not easily be explored before. Although big databases open up new opportunities for clinical research and data science, they also come with challenges. These include efficient data management as well as identification of biases and heterogeneity within the databases. This thesis aimed to address some of the challenges in the light of TBI analysis.

7.1 Summary of Findings

Data Analysis Pipelines. At first, two complementary pipelines were introduced for pre-processing structural or diffusion MRI scans. Both were designed with efficiency and flexibility in mind. The pipelines run independently for each subject, which allowed to process multiple scan sessions on a high performance computing cluster in parallel. The goal was a balance between accurate processing steps and minimised computation time. The structural pipeline first and foremost processes T1w images, however, also incorporates processing of other structural scans if available (e.g. FLAIR, T2w, SWI). For example, any provided complementary scan is coregistered to the T1w images, and additional feature maps are calculated (e.g. FLAIR^2). The diffusion pipeline allows to process DWI data with and without an extra b_0 image to apply EPI distortion correction. Furthermore, diffusion

parameter maps are coregistered to T1w space to make use of the T1w brain mask and allow multi-parametric data analysis that may require voxel-wise correspondence between T1w and diffusion images. Additionally, the pipeline automatically recognises multi-shell acquisitions and then applies free water elimination and kurtosis fitting. Besides extracting image derived features useful for clinical neuroscience, both pipelines provide several quality metrics which can be used to ensure the image usability and assess the accuracy of pre-processing steps. Quality control metrics were shown to pick up on scans with artefacts such as excessive head motion or noise corrupted images.

Application to Mild TBI. Both pipelines were employed to analyse mTBI data in a retrospective multi-centre study including three imaging sites. The heterogeneity of the data posed both possibilities and challenges. Scanning patients with varying severity of mTBI internationally bears the potential for the pooled database to reflect the spectrum of mTBI more accurately. However, differences in data collection scheme (e.g. scan time point after injury) and acquisition parameters hampered a straight-forward analysis. Despite rectifying data differences (e.g. using single-shell data exclusively) and applying the same processing pipeline, site-specific biases remained. To account for this, data were Z-scored and the acquisition site was used as covariate in the regression analysis. The focus laid on regional loss of brain tissue and changes in diffusion patterns. A comparison of control subjects and patients with different outcome revealed that patients with good outcome showed similar IDPs as control subjects. In contrast, patients with poor outcome experienced tissue atrophy, demonstrated as enlarged ventricles, and decreased FA indicating impaired WM integrity. Mean diffusivity was less sensitive to differences between controls and patient groups. Neither anatomical brain volumes nor diffusion parameters were found to be predictive of patient outcome. This suggested that injuries may not deviate strongly enough to separate between patient cohorts during the acute phase. The longitudinal analysis showed progressive tissue atrophy and changes of diffusion, suggesting IDPs may vary more between patient groups at a more chronic stage. However, there was no clear indicator that patients with poor outcome had stronger tissue atrophy or quicker deterioration of WM integrity. Predicting the outcome of mTBI patients remains a difficult problem, and it might be easier to assess how *control-like* the imaging features of a patient are.

Reproducibility of MRI Metrics. Combining different databases appropriately is challenging. To foster multi-centre studies it is important to understand the MR image reproducibility and its dependency on factors such as different acquisition parameters or scanners employed. For adequate analysis, the processing tools chosen to extract image features need

to be as robust to site-specific biases as possible. The analysis of whole brain parcellation (MALP-EM) revealed regional volume differences for the same subjects imaged on different scanners. The deviations could me minimised by standardising volumes to total brain volumes or partly by normalising image intensity prior to the ROI parcellation. Generally, smaller regions were more challenging to segment and showed higher variation when scans were acquired on different scanners or with different MR parameters. A registration based parcellation of WM (JHU) was found to be more robust to different MRI parameters than a model based ROI segmentation (TractSeg). However, registration was observed to be less precise in defining region boundaries. Besides ROI volumes, diffusion metrics were also found to be dependent on the acquisition scheme. Applying a different number of gradients of different strength and duration (single or multiple b-values) had an impact on both FA and MD. In particular, single-shell acquisition showed more variation in diffusion parameters than scans acquired on several shells. Overall, findings suggest that more similar parameters lead to less variation. Nonetheless, even with closely matched protocols image quality remains dependent on hardware settings. Pooling data from multiple different sites and scanners showed that variation did not exceed the highest variation found in individual databases.

Harmonisation of DWI for Multi-Centre Studies. With improved data available and the development of data-hungry algorithms, studies aim to increase sample sizes. Despite the effort of designing large prospective multi-centre studies, there is a demand to harmonise data across acquisition sites. Some promising methods were examined in depth to better understand their potential under a fair comparison. The previously reported success of linear scaling of SH coefficients via RISH feature mapping could only partly be reproduced in the experiments presented here. In particular higher order SH coefficients could not easily be mapped between scanners. Enhancements of scaling maps had only minor positive impact. In contrast, neural networks were shown to be more beneficial to learn non-local mappings between diffusion data to reduce site-specific biases. However, model based approaches were less robust to outliers than, for example, linear RISH feature scaling. The experiments also showed that neural networks could indeed profit from learning different feature maps on the various SH images of different orders. A caveat was that the designed neural networks were trained to learn a mapping between coregistered scans. Coregistering images resulted in decreased FA values for the data used for the experiments presented here. Consequently, the neural network learned to shift data to the *wrong* image domain (i.e. the coregistered images rather than the images in native space), which is why neural networks underperformed. Despite showing some potential to minimise site-specific variations, a benefit for

application to TBI patient data could not be observed.

Lesions Analysis in Severe TBI. Besides the analysis of mTBI, patients with more severe injuries were examined as well. One important factor in TBI is the presence and location of lesions visible on MR images. In the available dataset, FLAIR contusion cores and oedema were mostly found in the frontal and temporal lobes. Comparing two cohorts imaged on different scanners showed that contusions were predominately in opposite brain hemispheres. Although this is unrelated to the scanners employed, it highlighted potential variations in different cohorts across centres. The longitudinal data revealed that volumes for contusion cores initially grew in the acute phase, but seemed to shrink or vanish later on. In contrast, oedema volumes consistently increased with time post-injury. The algorithm introduced to match lesions between initial and follow up scans was successful for larger lesions. Therefore, measuring the lesion development of individual clusters has the potential to be more informative than simply assessing the total lesion volume.

7.2 Limitations & Future Directions

Data Analysis Pipelines. An obvious change for the pipeline will be to exchange tools for newer methods to keep the processing pipeline state of the art. This will involve the integration of improved methods such as, for example, reducing noise in diffusion weighted images. To streamline the pipelines even more, some modules will need to be replaced to accelerate processing steps. For example, instead of relying on registration (MALP-EM), the brain parcellation could be predicted by machine learning models. With respect to TBI, an important new extension of the pipeline will be the automated lesion segmentation and derivation of clinically relevant information such as the lesion location. Another extension will be the use of machine learning models trained on QC metrics to identify corrupted scans and processing steps. This will be more beneficial than relying on finding empirical thresholds for each particular QC metric to flag suspicious scans. Apart from that, the usability could be improved. The current state of the pipeline is easily applicable via command line, however, this might be challenging for the less trained clinical researcher. A better interface could strongly improve usability. A challenge will be to anticipate as many different scenarios as possible. The more flexible the pipeline will be, the higher the chance data will be processed inadequately.

Application to Mild TBI. Appropriate study design and processing becomes even more important if data from different databases are used. The pipelines were flexible enough to

allow their application to three different mTBI databases. Processing steps were mostly equally, but site-specific biases remained. Although these were also accounted for via Z-scoring and linear regression modelling, it is not entirely clear whether biases could be fully eliminated. For example, the higher shell diffusion data from Trondheim were disregarded which likely had a positive effect on the comparability of MD across centres. However, angular resolution - that rather influences FA than MD - was different across sites. This may have left FA more vulnerable to site-specific biases. These could be helpful in differentiating subjects based on their origin of acquisition, since the cohorts from the sites examined had different severities of injury (Trondheim patients had experienced milder TBI than Turku patients). Indeed, FA showed a higher sensitivity to differentiate controls and patient groups than MD. Future investigations will need to focus on better harmonisation strategies as well as being more patient-specific to combat the heterogeneity in imaging data and TBI cohorts. Furthermore, the data could be enhanced in two folds. Either increasing the sample size, by for example including CENTER-TBI data, or by exploiting multi-shell acquisition for better diffusion modelling (i.e. DKI and free-water elimination). The latter would need to rely on Trondheim data only.

Reproducibility of MRI Metrics. Experiments have shown the influence of MRI acquisition on image derived features (volumes and diffusion). Any study that pools data acquired on different scanners and/or with different parameters will be affected by site-specific biases, and analysis will need to account for that. Various datasets with different sources of variation were examined to understand their impact on image derived features. However, influences of various parameters were still very much entangled, and more experiments are needed to better understand the effect of single parameters. Although different acquisition schemes affected WM parcellation, a bigger challenge was the segmentation of WM tracts on Philips scanners. Since TractSeg is a predictive model, it drastically under-performed on unseen data, such as the Philips scans from CENTER-TBI. Tract segmentation via registration (JHU) tract segmentation was more robust but less precise. Therefore, future work will need to focus on multi-centre WM parcellation to allow a reliable analysis of the whole CENTER-TBI database. This could involve either data harmonisation prior to the parcellation or designing models that learn to adapt to different domains.

Harmonisation of DWI for Multi-Centre Studies. Diffusion MRI harmonisation remains an open research question. One of the limitations in the experiments presented here was the low sample size. Both the linear RISH feature scaling and neural networks would benefit from including more data. However, acquiring data for travelling healthy

controls is expensive and highly unpractical for large multi-centre studies. Even collecting MRI data for enough matched (e.g. age, sex) control subjects is challenging. With respect to CENTER-TBI, only a maximum nine healthy volunteers were scanned per site. Since most of these volunteers are different for each scanner, matching controls will be difficult. Furthermore, a small number of control subjects may not reflect the full data distribution, which could skew any analysis following the harmonisation. Therefore, new methods will need to be developed that can be applied to small datasets of possibly unmatched controls. Furthermore, the methods investigated here all aim to project one database to another. This is infeasible for multi-centre studies as one mapping for each individual database would need to be learned which makes data harmonisation less robust. A key development will be simultaneous representation learning of diffusion MRI, such that the information of all data can be leveraged at once. Possibly this would also allow the inclusion of patient data in the learning process. Instead of learning a mapping between control subjects to harmonise patient data before analysis, a strong model could learn to reduce site-specific variation while simultaneously finding disease related differences. This could especially be advantageous for studies such as CENTER-TBI where control subjects are limited. In addition, non-imaging characteristics could be directly incorporated during the learning phase. One challenge, however, will be the interpretability of the neural network's decisions to infer clinically relevant information. Another research direction could focus on disentangling the different effects that lead to site-related biases. If the influences of different acquisition parameters could be differentiated from hardware induced effects, control subjects from another study, imaged with different MRI protocol but on the same scanner, could help to learn site-specific biases. Disentanglement learning is, however, a fairly new research field.

Lesions Analysis in Severe TBI. Lesion analysis is an important factor to examine severe TBI cohorts. Two algorithms were introduced to foster automated examination of FLAIR lesions. The region-based localisation of lesion is highly dependent on the atlas employed. If the atlas is too crude, it is difficult to actually locate lesions. For example, the MALP-EM atlas does not parcellate WM regions. With WM as the largest region, almost any lesion was located within WM, which makes differentiation of lesions by location more difficult. On the other hand if the parcellation is too fine, a lesion will be present in many small regions, which may hamper the identification of a distinct lesion location. Future experiments will need to investigate the use of different atlases and the automated extraction of addition quantifiable measures (for example sphericity or surface area of a lesion). If lesions are more prone to grow than others, therapy could focus on those. Therefore, besides assessing the momentary characteristics of a lesion on one scan, it is important to

understand the lesion evolution over time. The lesion matching algorithm was designed to connect individual lesion clusters between initial and follow-up scan. Since this algorithm is based on registration and overlap of lesion annotations, it is prone to fail to match small lesions. This can lead to spurious results as small lesions may appear to be salvageable or appear at a later stage. More experiments on a larger database are needed to understand the full potential. Once lesions will be segmented automatically for the full CENTER-TBI database, these algorithms will be applicable to investigate the connection between extracted lesion characteristics and clinical variables (e.g. outcome or mortality) on a larger scale.

7.3 Conclusion

Neuroimaging studies are increasing in size. While this can beneficial for data analysis to boost statistical power and examine bigger patient cohorts, this comes also with challenges. Databases need to be managed and processed in an efficient and transparent way that fosters reproducible data analysis. Flexible processing pipeline that robustly remove artefact and extract image derived features, while being adaptive to the needs of different databases, support answering clinical neuroscience questions. These are particularly complex for heterogeneous databases such as TBI patient cohorts. Depending on the severity of the injury, patients show diffuse axonal injury best observed on diffusion MRI, or clearly visible lesions. This high variability in pathology requires different analysis strategies. While the challenge for mTBI cases lies in detecting very subtle differences between patients, severe TBI cases may benefit from extraction of lesion characteristics. Generally, collecting big neuroimaging data is challenging which is why there is demand to merge databases across sites. This can happen retrospectively for independent studies, or as part of a prospectively designed multi-centre study. Either way, pooling data from different centres comes with increased variation. Acquisition dependent biases are inherently introduced and need to be accounted for. Especially for diffusion MRI for which angular resolution and the strength and duration of the applied gradients strongly influence the measured diffusion signal. Newest developments in MRI data harmonisation showed the great success of deep learning neural networks to minimise site-specific biases, however, more experiments are needed to fully understand the potential for clinical analysis of TBI data. Big databases, such as CENTER-TBI, and advanced algorithms make TBI research an exiting field that still holds many challenges. The hope is that larger datasets, more consistent pre-processing tools and improved data harmonisation methods will help to analyse TBI cohorts to ultimately lead to revelations that improve patient care.

Appendix A

Table A.1: P-Values After FDR Correction for Comparison of TractSeg ROI Mean FA Within and Across Scanners. P-values of post-hoc t-test <0.05 printed in bold.

		Intra-Scanner		Inter-Scanner			
ROI	rm-ANOVA	P1-P2	T1-T2	P1-T1	P1-T1	P2-T1	P2-T2
FA							
14	0.0060	0.4021	0.9177	0.0049	0.0076	0.0613	0.0720
17	0.0157	0.0419	0.3799	0.0419	0.0640	0.0776	0.2369
18	0.0094	0.1580	0.9113	0.0058	0.0248	0.0341	0.1157
28	0.0157	0.0077	0.8507	0.0521	0.0236	0.1255	0.0671
31	0.0405	0.2440	0.9996	0.0132	0.0148	0.2440	0.2440
32	0.0116	0.0133	0.9658	0.0156	0.0156	0.1395	0.2797
33	0.0001	0.0941	0.9065	0.0000	0.0006	0.0182	0.0286
41	0.0000	0.1965	0.2454	0.0033	0.0007	0.0078	0.0016
45	0.0157	0.0405	0.9467	0.0828	0.1073	0.0405	0.0562
50	0.0157	0.2874	0.2874	0.0536	0.0536	0.1488	0.1003
52	0.0129	0.3720	0.8004	0.0197	0.0189	0.0829	0.0829
58	0.0040	0.1873	0.8554	0.0414	0.0128	0.0251	0.0128
64	0.0321	0.2533	0.2533	0.0588	0.0588	0.2303	0.1578
66	0.0163	0.4278	0.5731	0.0274	0.0155	0.1443	0.1198

ROIs: 0: left arcuate fascicle, 6: genu, 7: rostral body of CC, 22: left inferior cerebellar peduncle, 25: right IFO, 30: right optic radiation, 31: left parieto-occipital pontine, 32: right parieto-occipital pontine, 36: right SLF_I, 45: CC, 48: left thalamo-premotor tract, 57: right thalamo-occipital tract, 71: right striato-occipital tract. **Scans:** P1: Prisma scan #1, P2: Prisma scan #2, T1: Trio scan #1, T2: Trio scan #2.

Table A.2: P-Values After FDR Correction for Comparison of TractSeg ROI Mean MD Within and Across Scanners. P-values of post-hoc t-test <0.05 printed in bold.

		Intra-Scanner		Inter-Scanner			
ROI	rm-ANOVA	P1-P2	T1-T2	P1-T1	P1-T1	P2-T1	P2-T2
MD							
5	0.0333	0.2757	0.5226	0.0998	0.1502	0.0998	0.0998
6	0.0149	0.4974	0.1078	0.0733	0.0733	0.0768	0.0768
9	0.0395	0.4232	0.4232	0.1560	0.1560	0.1560	0.1560
10	0.0164	0.3067	0.4024	0.0783	0.1968	0.0537	0.1289
14	0.0131	0.3854	0.7387	0.0057	0.0218	0.1268	0.1268
18	0.0164	0.3438	0.8842	0.0081	0.0081	0.2077	0.2077
20	0.0186	0.2993	0.0279	0.7648	0.1002	0.1284	0.0279
21	0.0163	0.6538	0.2363	0.1361	0.0660	0.0707	0.0583
22	0.0117	0.6179	0.4133	0.0190	0.0190	0.1356	0.0190
23	0.0027	0.4060	0.4060	0.0211	0.0131	0.0131	0.0131
24	0.0161	0.5126	0.2198	0.0825	0.1338	0.0519	0.0825
26	0.0159	0.5372	0.1369	0.0713	0.0454	0.1063	0.0282
27	0.0285	0.3001	0.1370	0.1370	0.1370	0.1370	0.1370
28	0.0004	0.6087	0.6087	0.0057	0.0040	0.0205	0.0057
30	0.0200	0.1145	0.1472	0.1159	0.1145	0.1159	0.1145
33	0.0016	0.4546	0.7146	0.0020	0.0021	0.0446	0.0261
34	0.0004	0.2149	0.7085	0.0044	0.0048	0.0054	0.0248
36	0.0094	0.7743	0.2857	0.0438	0.1112	0.0300	0.0678
41	0.0006	0.1460	0.4354	0.0016	0.0062	0.0323	0.0323
45	0.0015	0.3929	0.1415	0.0196	0.0394	0.0196	0.0394
51	0.0103	0.1159	0.4680	0.0378	0.1159	0.0127	0.0378
53	0.0004	0.1016	0.5783	0.0013	0.0194	0.0008	0.0167
54	0.0347	0.4286	0.4705	0.0611	0.2813	0.0611	0.1962
55	0.0175	0.3381	0.4923	0.0988	0.1888	0.0609	0.1198
57	0.0163	0.5059	0.1373	0.1249	0.1040	0.1363	0.1040
65	0.0044	0.1344	0.3487	0.0297	0.1200	0.0074	0.0297
67	0.0004	0.1711	0.4556	0.0007	0.0157	0.0001	0.0131
68	0.0175	0.3594	0.3594	0.0564	0.1880	0.0564	0.1419
69	0.0164	0.2132	0.4434	0.1025	0.1856	0.0552	0.1025
71	0.0175	0.1426	0.4318	0.0738	0.0738	0.0738	0.0738

ROIs: 0: left arcuate fascicle, 6: genu of corpus callosum, 7: rostral body of corpus callosum, 22: left inferior cerebellar peduncle, 25: right IFO fascicle, 30: right optic radiation, 31: left parieto-occipital pontine, 32: right parieto-occipital pontine, 36: right superior longitudinal fasicle I, 45: full corpus callosum, 48: left thalamo-premotor, 57: right thalamo-occipital, 71: right striato-occipital. **Scans:** P1: Prisma scan #1, P2: Prisma scan #2, T1: Trio scan #1, T2: Trio scan #2.

Table A.3: P-Values After FDR Correction for Comparison of JHU ROI Mean DTI Metrics Within and Across Scanners. P-values of post-hoc t-test <0.05 printed in bold.

		Intra-Scanner		Inter-Scanner			
ROI	rm-ANOVA	P1-P2	T1-T2	P1-T1	P1-T1	P2-T1	P2-T2
FA							
0	0.0327	0.1230	0.7117	0.0944	0.1148	0.1837	0.2866
1	0.0002	0.0413	0.7112	0.0042	0.0042	0.0413	0.0065
2	0.0001	0.2382	0.2382	0.0017	0.0021	0.0017	0.0132
6	0.0032	0.9136	0.0662	0.0085	0.0662	0.0085	0.1444
9	0.0151	0.7641	0.9010	0.0889	0.0550	0.0550	0.0051
14	0.0026	0.2890	0.8990	0.0043	0.0098	0.0858	0.0679
18	0.0471	0.3484	0.3484	0.0352	0.2030	0.2482	0.2616
MD							
0	0.0011	0.1790	0.8637	0.0202	0.0202	0.0202	0.0299
1	0.0021	0.2882	0.8223	0.0207	0.0209	0.0207	0.0455
2	0.0007	0.5181	0.7019	0.0048	0.0048	0.0048	0.0128
4	0.0255	0.4138	0.5224	0.1033	0.1033	0.1182	0.1381
6	0.0031	0.5208	0.7153	0.0389	0.0094	0.0971	0.0161
7	0.0226	0.9535	0.0050	0.2173	0.0728	0.2156	0.0728
12	0.0022	0.1880	0.4423	0.0270	0.0270	0.0270	0.0270
13	0.0031	0.5581	0.3812	0.0763	0.0344	0.0763	0.0344
14	0.0200	0.4770	0.4992	0.0031	0.0405	0.2046	0.2046
18	0.0022	0.4397	0.3083	0.0007	0.0048	0.0806	0.0698

ROIs: 0: left anterior thalamic radiation, 1: right anterior thalamic radiation, 2: left CST, 4: left CG, 6: left hypocampal cingulate, 7: right hypocampal cingulate, 9: forceps minor,12: left ILF, 13: right ILF, 14: left SLF, 18: left SLF. **Scans:** P1: Prisma scan #1, P2: Prisma scan #2, T1: Trio scan #1, T2: Trio scan #2.

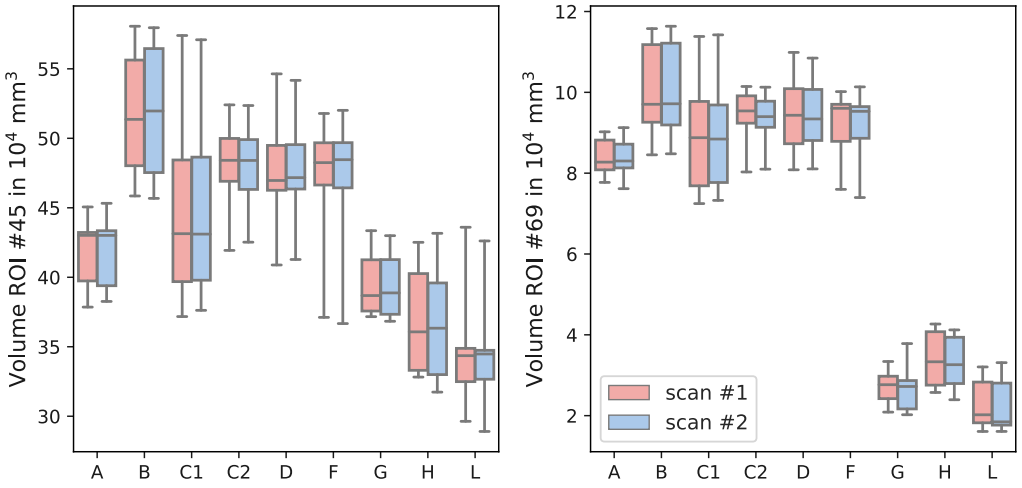


Figure A.1: TractSeg Volume Distribution for CENTER-TBI. Volume distributions in different CENTER-TBI imaging sites. **Left:** The corpus callosum (ROI #45) as largest segmented region in the TractSeg atlas was segmented for in all centres, however, much lower volumes on Philips scanners (G, H, L) were observed. **Right:** Volumes of the left striato-occipital tract (ROI #69), measured on Philips scans, were strongly undersegmented in comparison to all other centres that employed a different vendor (i.e. GE or Siemens). This region is representative for many other undersegmented ROIs on Philips data.

Table 7.4: Effect Size of Harmonisation Methods on Control Subjects from CENTER-TBI. Cohen's effect size d was computed between global RMSE values. Effect sizes can be small ($d = \pm 0.2$), medium ($d = \pm 0.5$) or large ($d = \pm 0.8$).

Method #1	Method #2	b_0	$RISH_0$	$RISH_2$	$RISH_4$
Inter-Scanner	Linear-RISH	0.97	1.14	1.40	-0.31
Inter-Scanner	Global Scaling	0.89	1.17	1.22	-1.08
Inter-Scanner	CNN Multi-Path	1.38	1.37	1.73	1.61
Linear-RISH	Global Scaling	0.10	0.06	-0.37	-0.95
Linear-RISH	CNN Multi-Path	1.03	0.55	1.18	1.72
Global Scaling	CNN Multi-Path	1.03	0.49	1.31	1.75

Cohen's effect size measure: $d = (\overline{x} - \overline{y})/\sigma$, with $\overline{x}$ and $\overline{y}$ representing the mean of the RMSE (between the corresponding images from both scanners) for method #1 and method #2, respectively. σ represents the pooled standard deviation of the RMSE of both methods.

Bibliography

[1] Krishma Adatia, Virginia FJ Newcombe, and David K Menon. Contusion progression following traumatic brain injury: a review of clinical and radiological predictors, and influence on outcome. *Neurocritical care*, pages 1–13, 2020.

[2] Mohamed N Ahmed, Sameh M Yamany, Nevin Mohamed, Aly A Farag, and Thomas Moriarty. A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *IEEE transactions on medical imaging*, 21(3):193–199, 2002.

[3] Andrew L Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.

[4] Naomi Allen, Cathie Sudlow, Paul Downey, Tim Peakman, John Danesh, Paul Elliott, John Gallacher, Jane Green, Paul Matthews, Jill Pell, et al. Uk biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126, 2012.

[5] Sarah M Andersen, Steven Z Rapcsak, and Pélagie M Beeson. Cost function masking during normalization of brains with focal lesions: still a necessity? *Neuroimage*, 53(1):78–84, 2010.

[6] Jesper LR Andersson, Stefan Skare, and John Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, 20(2):870–888, 2003.

[7] Jesper LR Andersson and Stamatios N Sotiropoulos. Non-parametric representation and prediction of single-and multi-shell diffusion-weighted mri data using gaussian processes. *Neuroimage*, 122:166–176, 2015.

[8] Jesper LR Andersson and Stamatios N Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging. *Neuroimage*, 125:1063–1078, 2016.

[9] Yuta Aoki, Ryota Inokuchi, Masataka Gunshin, Naoki Yahagi, and Hiroshi Suwa. Diffusion tensor imaging studies of mild traumatic brain injury: a meta-analysis. *J Neurol Neurosurg Psychiatry*, 83(9):870–876, 2012.

[10] John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000.

[11] Stephen Ashwal, Karen A Tong, Nirmalya Ghosh, Brenda Bartnik-Olson, and Barbara A Holshouser. Application of advanced neuroimaging modalities in pediatric traumatic brain injury. *Journal of child neurology*, 29(12):1704–1717, 2014.

[12] Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight j*, 2:1–35, 2009.

[13] M Sunil Babu and V Vijayalakshmi. A review on acute/sub-acute ischemic stroke lesion segmentation and registration challenges. *Multimedia Tools and Applications*, 78(2):2481–2506, 2019.

[14] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 2019.

[15] Gonzalo Barrio-Arranz, Rodrigo de Luis-García, Antonio Tristán-Vega, Marcos Martín-Fernández, and Santiago Aja-Fernández. Impact of mr acquisition parameters on dti scalar indexes: a tractography based approach. *PloS one*, 10(10):e0137905, 2015.

[16] Peter J Basser and Carlo Pierpaoli. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri. *Journal of magnetic resonance*, 213(2):560–570, 2011.

[17] Matteo Bastiani, Michiel Cottaar, Sean P Fitzgibbon, Sana Suri, Fidel Alfaro-Almagro, Stamatios N Sotiropoulos, Saad Jbabdi, and Jesper LR Andersson. Automated quality control for within and between studies diffusion mri data using a non-parametric framework for movement and distortion correction. *NeuroImage*, 184:801–812, 2019.

[18] Carrie E Bearden and Paul M Thompson. Emerging global initiatives in neurogenetics: the enhancing neuroimaging genetics through meta-analysis (enigma) consortium. *Neuron*, 94(2):232–236, 2017.

[19] Barbara B Bendlin, Michele L Ries, Mariana Lazar, Andrew L Alexander, Robert J Dempsey, Howard A Rowley, Jack E Sherman, and Sterling C Johnson. Longitudinal changes in patients with traumatic brain injury assessed with diffusion-tensor and volumetric imaging. *Neuroimage*, 42(2):503–514, 2008.

[20] Erin D Bigler. Traumatic brain injury, neuroimaging, and neurodegeneration. *Frontiers in human neuroscience*, 7:395, 2013.

[21] Erin D Bigler, Tracy J Abildskov, JoAnn Petrie, Thomas J Farrer, Maureen Dennis, Nevena Simic, H Gerry Taylor, Kenneth H Rubin, Kathryn Vannatta, Cynthia A Gerhardt, et al.

Heterogeneity of brain lesions in pediatric traumatic brain injury. *Neuropsychology*, 27(4):438, 2013.

[22] Erin D Bigler and Jeffrey J Bazarian. Diffusion tensor imaging: a biomarker for mild traumatic brain injury? *Neurology*, 74(8):626–627, 2010.

[23] Erin D Bigler, Stephen R McCauley, Trevor C Wu, Ragini Yallampalli, Sanjeev Shah, Marianne MacLeod, Zili Chu, Jill V Hunter, Guy L Clifton, Harvey S Levin, et al. The temporal stem in traumatic brain injury: preliminary findings. *Brain imaging and behavior*, 4(3):270–282, 2010.

[24] S Bisdas, DE Bohning, N Bešenski, JS Nicholas, and Z Rumboldt. Reproducibility, interrater agreement, and age-related changes of fractional anisotropy measures at 3t in healthy subjects: effect of the applied b-value. *American Journal of Neuroradiology*, 29(6):1128–1133, 2008.

[25] Helen M Bramlett and W Dalton Dietrich. Long-term consequences of traumatic brain injury: current status of potential mechanisms of injury and neurological outcomes. *Journal of neurotrauma*, 32(23):1834–1848, 2015.

[26] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] Matthew Brett, Alexander P Leff, Chris Rorden, and John Ashburner. Spatial normalization of brain images with focal lesions using cost function masking. *Neuroimage*, 14(2):486–500, 2001.

[28] Adrian Burton. The center-tbi core study: The making-of. *The Lancet Neurology*, 16(12):958–959, 2017.

[29] Joseph A Carnevale, David J Segar, Andrew Y Powers, Meghal Shah, Cody Doberstein, Benjamin Drapcho, John F Morrison, John R Williams, Scott Collins, Kristina Monteiro, et al. Blossoming contusions: identifying factors contributing to the expansion of traumatic intracerebral hemorrhage. *Journal of neurosurgery*, 1(aop):1–12, 2018.

[30] Daniel C Castro and Ben Glocker. Nonparametric density flows for mri intensity normalisation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 206–214. Springer, 2018.

[31] Mara Cercignani, Roland Bammer, Maria P Sormani, Franz Fazekas, and Massimo Filippi. Inter-sequence and inter-imaging unit variability of diffusion tensor mr imaging histogram-derived metrics of the brain in healthy volunteers. *American journal of neuroradiology*, 24(4):638–643, 2003.

[32] Suheyla Cetin-Karayumak, Maria A Di Biase, Natalia Chunga, Benjamin Reid, Nathaniel Somes, Amanda E Lyall, Sinead Kelly, Bengisu Solgun, Ofer Pasternak, Mark Vangel, et al. White matter abnormalities across the lifespan of schizophrenia: a harmonized multi-site diffusion mri study. *Molecular psychiatry*, pages 1–12, 2019.

[33] Suheyla Cetin-Karayumak, Katharina Stegmayer, Sebastian Walther, Philip R Szeszko, Tim Crow, Anthony James, Matcheri Keshavan, Marek Kubicki, and Yogesh Rathi. Exploring the limits of combat method for multi-site diffusion mri harmonization. *bioRxiv*, 2020.

[34] Maxime Chamberland, Sila Genc, Erika P Raven, Greg D Parker, Adam Cunningham, Joanne Doherty, Marianne van den Bree, Chantal MW Tax, and Derek K Jones. Tractometry-based anomaly detection for single-subject white matter analysis. *arXiv preprint arXiv:2005.11082*, 2020.

[35] Cody A Chastain, Udochukwu E Oyoyo, Michelle Zipperman, Elliot Joo, Stephen Ashwal, Lori A Shutter, and Karen A Tong. Predicting outcomes of traumatic brain injury by imaging modality and injury distribution. *Journal of neurotrauma*, 26(8):1183–1196, 2009.

[36] Andrew A Chen, Joanne C Beer, Nicholas J Tustison, Philip A Cook, Russell T Shinohara, Haochang Shou, Alzheimer's Disease Neuroimaging Initiative, et al. Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. *bioRxiv*, page 858415, 2019.

[37] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[38] James H Cole, Amy Jolly, Sara de Simoni, Niall Bourke, Maneesh C Patel, Gregory Scott, and David J Sharp. Spatial patterns of progressive brain volume loss after moderate-severe traumatic brain injury. *Brain*, 141(3):822–836, 2018.

[39] James H Cole, Robert Leech, David J Sharp, and Alzheimer's Disease Neuroimaging Initiative. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*, 77(4):571–581, 2015.

[40] Kara N Corps, Theodore L Roth, and Dorian B McGavern. Inflammation and neuroprotection in traumatic brain injury. *JAMA neurology*, 72(3):355–362, 2015.

[41] Marta Morgado Correia, Thomas A Carpenter, and Guy B Williams. Looking for the optimal dti acquisition scheme given a maximum scan time: are more b-values a waste of time? *Magnetic resonance imaging*, 27(2):163–175, 2009.

[42] Jennifer T Crinion, Matthew A Lambon-Ralph, Elizabeth A Warburton, David Howard, and Richard JS Wise. Temporal lobe regions engaged during normal speech comprehension. *Brain*, 126(5):1193–1201, 2003.

[43] Iain D Croall, Christopher JA Cowie, Jiabao He, Anna Peel, Joshua Wood, Benjamin S Aribisala, Patrick Mitchell, A David Mendelow, Fiona E Smith, David Millar, et al. White matter correlates of cognitive dysfunction after mild traumatic brain injury. *Neurology*, 83(6):494–501, 2014.

[44] Leo F Czervionke, Jeanne M Czervionke, David L Daniels, and Victor M Haughton. Characteristic features of mr truncation artifacts. *American journal of roentgenology*, 151(6):1219–1228, 1988.

[45] Daniel H Daneshvar, David O Riley, Christopher J Nowinski, Ann C McKee, Robert A Stern, and Robert C Cantu. Long-term consequences: effects on normal development profile after concussion. *Physical Medicine and Rehabilitation Clinics*, 22(4):683–700, 2011.

[46] Sandhitsu R Das, Brian B Avants, Murray Grossman, and James C Gee. Registration based cortical thickness measurement. *Neuroimage*, 45(3):867–879, 2009.

[47] Flavio Dell'Acqua, Luis Lacerda, Marco Catani, and Andrew Simmons. Anisotropic power maps: A diffusion contrast to reveal low anisotropy tissues from hardi data. In *Proceedings Joint Annual Meeting ISMRMESMRMB, ISMRM2014, Milan*, page 0730, 2014.

[48] Aurélie Delouche, Arnaud Attyé, Olivier Heck, Sylvie Grand, Adrian Kastler, Laurent Lamalle, Felix Renard, and Alexandre Krainik. Diffusion mri: pitfalls, literature review and future directions of research in mild traumatic brain injury. *European journal of radiology*, 85(1):25–30, 2016.

[49] S Deprez, Michiel B de Ruiter, S Bogaert, R Peeters, J Belderbos, D De Ruysscher, S Schagen, S Sunaert, P Pullens, and E Achten. Multi-center reproducibility of structural, diffusion tensor, and resting state functional magnetic resonance imaging measures. *Neuroradiology*, 60(6):617–634, 2018.

[50] Maxime Descoteaux. High angular resolution diffusion imaging (hardi). *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–25, 1999.

[51] Michael C Dewan, Abbas Rattani, Saksham Gupta, Ronnie E Baticulon, Ya-Ching Hung, Maria Punchak, Amit Agrawal, Amos O Adeleye, Mark G Shrime, Andrés M Rubiano, et al. Estimating the global incidence of traumatic brain injury. *Journal of neurosurgery*, 130(4):1080–1097, 2018.

[52] Raunak Dey and Yi Hong. Compnet: Complementary segmentation network for brain mri extraction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 628–636. Springer, 2018.

[53] Thijs Dhollander, David Raffelt, and Alan Connelly. Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion mr data without a co-registered t1 image. In *ISMRM Workshop on Breaking the Barriers of Diffusion MRI*, volume 5, 2016.

[54] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[55] Olaf Dietrich, José G Raya, Scott B Reeder, Michael Ingrisch, Maximilian F Reiser, and Stefan O Schoenberg. Influence of multichannel combination, parallel imaging and other reconstruction techniques on mri noise characteristics. *Magnetic resonance imaging*, 26(6):754–762, 2008.

[56] James J Donkin and Robert Vink. Mechanisms of cerebral edema in traumatic brain injury: therapeutic developments. *Current opinion in neurology*, 23(3):293–299, 2010.

[57] Reuben Dorent, Wenqi Li, Jinendra Ekanayake, Sebastien Ourselin, and Tom Vercauteren. Learning joint lesion and tissue segmentation from task-specific hetero-modal datasets. In *International Conference on Medical Imaging with Deep Learning*, pages 164–174, 2019.

[58] David B Douglas, Tae Ro, Thomas Toffoli, Bennet Krawchuk, Jonathan Muldermans, James Gullo, Adam Dulberger, Ariana E Anderson, Pamela K Douglas, and Max Wintermark. Neuroimaging of traumatic brain injury. *Medical Sciences*, 7(1):2, 2019.

[59] Robert R Edelman. The history of mr imaging as seen through the pages of radiology. *Radiology*, 273(2S):S181–S200, 2014.

[60] Brian L Edlow, William A Copen, Saef Izzy, Khamid Bakhadirov, Andre van der Kouwe, Mel B Glenn, Steven M Greenberg, David M Greer, and Ona Wu. Diffusion tensor imaging in acute-to-subacute traumatic brain injury: a longitudinal analysis. *BMC neurology*, 16(1):2, 2016.

[61] Amaal Eman Abdulle and Joukje van der Naalt. The role of mood, post-traumatic stress, post-concussive symptoms and coping on outcome after mtbi in elderly patients. *International Review of Psychiatry*, pages 1–9, 2019.

[62] Sophie Eyres, Amy Carey, Gill Gilworth, Vera Neumann, and Alan Tennant. Construct validity and reliability of the rivermead post-concussion symptoms questionnaire. *Clinical rehabilitation*, 19(8):878–887, 2005.

[63] Irene Fantini, Leticia Rittner, Clarissa Yasuda, and Roberto Lotufo. Automatic detection of motion artifacts on mri using deep cnn. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE, 2018.

[64] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[65] Moshiur R Farazi, Fahim Faisal, Zaied Zaman, and Soumik Farhan. Inpainting multiple sclerosis lesions for improving registration performance with brain atlas. In *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*, pages 1–6. IEEE, 2016.

[66] Jonathan AD Farrell, Bennett A Landman, Craig K Jones, Seth A Smith, Jerry L Prince, Peter CM Van Zijl, and Susumu Mori. Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging–derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 t. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 26(3):756–767, 2007.

[67] Anders M Fjell, Kristine B Walhovd, Christine Fennema-Notestine, Linda K McEvoy, Donald J Hagler, Dominic Holland, James B Brewer, and Anders M Dale. One-year brain atrophy evident in healthy aging. *Journal of Neuroscience*, 29(48):15223–15231, 2009.

[68] Centers for Disease Control, Prevention, et al. Report to congress on mild traumatic brain injury in the united states: steps to prevent a serious public health problem. *Atlanta, GA: Centers for Disease Control and Prevention*, 45, 2003.

[69] Jean-Philippe Fortin, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120, 2018.

[70] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, 2017.

[71] Jean-Philippe Fortin, Elizabeth M Sweeney, John Muschelli, Ciprian M Crainiceanu, Russell T Shinohara, Alzheimer's Disease Neuroimaging Initiative, et al. Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage*, 132:198–212, 2016.

[72] RJ Fox, K Sakaie, J-C Lee, JP Debbins, Y Liu, DL Arnold, ER Melhem, CH Smith, MD Philips, M Lowe, et al. A validation study of multicenter diffusion tensor imaging: reliability of fractional anisotropy and diffusivity values. *American journal of neuroradiology*, 33(4):695–700, 2012.

[73] Michael Gaetz. The neurophysiology of brain injury. *Clinical neurophysiology*, 115(1):4–18, 2004.

[74] Marco Ganzetti, Nicole Wenderoth, and Dante Mantini. Whole brain myelin mapping using t1-and t2-weighted mr imaging data. *Frontiers in human neuroscience*, 8:671, 2014.

[75] AJ Gardner and R Zafonte. Neuroepidemiology of traumatic brain injury. In *Handbook of clinical neurology*, volume 138, pages 207–223. Elsevier, 2016.

[76] Sairam Geethanath and John Thomas Vaughan Jr. Accessible magnetic resonance imaging: A review. *Journal of Magnetic Resonance Imaging*, 49(7):e65–e77, 2019.

[77] Allan George, Ruben Kuzniecky, Henry Rusinek, Heath R Pardoe, and Human Epilepsy Project Investigators. Standardized brain mri acquisition protocols improve statistical power in multicenter quantitative morphometry studies. *Journal of Neuroimaging*, 30(1):126–133, 2020.

[78] Bram HJ Geurts, Teuntje MJC Andriessen, Bozena M Goraj, and Pieter E Vos. The reliability of magnetic resonance imaging in traumatic brain injury lesion detection. *Brain injury*, 26(12):1439–1450, 2012.

[79] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[80] Marco Giannelli, Mirco Cosottini, Maria Chiara Michelassi, Guido Lazzarotti, Gina Belmonte, Carlo Bartolozzi, and Mauro Lazzeri. Dependence of brain dti maps of fractional anisotropy and mean diffusivity on the number of diffusion weighting directions. *Journal of applied clinical medical physics*, 11(1):176–190, 2010.

[81] Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

[82] Gary H Glover, Bryon A Mueller, Jessica A Turner, Theo GM Van Erp, Thomas T Liu, Douglas N Greve, James T Voyvodic, Jerod Rasmussen, Gregory G Brown, David B Keator, et al. Function biomedical informatics research network recommendations for prospective multicenter functional mri studies. *Journal of Magnetic Resonance Imaging*, 36(1):39–54, 2012.

[83] Lizzette Gómez-de Regil. Assessment of executive function in patients with traumatic brain injury with the wisconsin card-sorting test. *Brain Sciences*, 10(10):699, 2020.

[84] Rafael Gomez-Hernandez, Jeffrey E Max, Todd Kosier, Sergio Paradiso, and Robert G Robinson. Social impairment and depression after traumatic brain injury. *Archives of physical medicine and rehabilitation*, 78(12):1321–1326, 1997.

[85] Sandra González-Villà, Arnau Oliver, Yuankai Huo, Xavier Lladó, and Bennett A Landman. Brain structure segmentation in the presence of multiple sclerosis lesions. *NeuroImage: Clinical*, 22:101709, 2019.

[86] Krzysztof Gorgolewski, Christopher D Burns, Cindee Madison, Dav Clark, Yaroslav O Halchenko, Michael L Waskom, and Satrajit S Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5:13, 2011.

[87] Mark S Graham, Ivana Drobnjak, and Hui Zhang. A supervised learning approach for diffusion mri quality control with minimal training data. *NeuroImage*, 178:668–676, 2018.

[88] Neil SN Graham, Amy Jolly, Karl Zimmerman, Niall J Bourke, Gregory Scott, James H Cole, Jonathan M Schott, and David J Sharp. Diffuse axonal injury predicts neurodegeneration after moderate–severe traumatic brain injury. *Brain*, 143(12):3685–3698, 2020.

[89] Benjamin Y Gravesteijn, Daan Nieboer, Ari Ercole, Hester F Lingsma, David Nelson, Ben Van Calster, Ewout W Steyerberg, Cecilia Åkerlund, Krisztina Amrein, Nada Andelic, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of clinical epidemiology*, 2020.

[90] Hákon Gudbjartsson and Samuel Patz. The rician distribution of noisy mri data. *Magnetic resonance in medicine*, 34(6):910–914, 1995.

[91] Jared Hamwood, David Alonso-Caneiro, Scott A Read, Stephen J Vincent, and Michael J Collins. Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of oct retinal layers. *Biomedical optics express*, 9(7):3049–3066, 2018.

[92] Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, et al. Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, 32(1):180–194, 2006.

[93] Taylor C Harris, Rijk de Rooij, and Ellen Kuhl. The shrinking brain: cerebral atrophy following traumatic brain injury. *Annals of biomedical engineering*, 47(9):1941–1959, 2019.

[94] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016.

[95] Colin Hawco, Joseph D Viviano, Sofia Chavez, Erin W Dickie, Navona Calarco, Peter Kochunov, Miklos Argyelan, Jessica A Turner, Anil K Malhotra, Robert W Buchanan, et al. A longitudinal human phantom reliability study of multi-center t1-weighted, dti, and resting state fmri data. *Psychiatry Research: Neuroimaging*, 282:134–142, 2018.

[96] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[97] Eun Hee Kwak, Soohyun Wi, MinGi Kim, Soonil Pyo, Yoon-Kyum Shin, Kyung Ja Oh, Kyunghun Han, Yong Wook Kim, and Sung-Rae Cho. Factors affecting cognition and emotion in patients with traumatic brain injury. *NeuroRehabilitation*, (Preprint):1–11, 2020.

[98] Torgeir Hellstrøm, Lars Tjelta Westlye, Solrun Sigurdardottir, Cathrine Brunborg, HL Soberg, Øyvor Holthe, ANDRES Server, Martina Jonette Lund, Ole Andreas Andreassen, and Nada Andelic. Longitudinal changes in brain morphology from 4 weeks to 12 months after mild traumatic brain injury: associations with cognitive functions and clinical variables. *Brain injury*, 31(5):674–685, 2017.

[99] Guillaume Herbet, Ilyess Zemmoura, and Hugues Duffau. Functional anatomy of the inferior longitudinal fasciculus: from historical reports to current hypotheses. *Frontiers in neuroanatomy*, 12:77, 2018.

[100] Andrew R Hoy, Cheng Guan Koay, Steven R Kecskemeti, and Andrew L Alexander. Optimization of a free water elimination two-compartment model for diffusion tensor imaging. *Neuroimage*, 103:323–333, 2014.

[101] Thierry AGM Huisman, Thomas Loenneker, Gerd Barta, Matthias E Bellemann, Juergen Hennig, Joachim E Fischer, and Kamil A Il'yasov. Quantitative diffusion tensor mr imaging of the brain: field strength related variance of apparent diffusion coefficient (adc) and fractional anisotropy (fa) scalars. *European radiology*, 16(8):1651, 2006.

[102] MB Hulkower, DB Poliak, SB Rosenbaum, ME Zimmerman, and Michael L Lipton. A decade of dti in traumatic brain injury: 10 years and 100 articles later. *American Journal of Neuroradiology*, 34(11):2064–2074, 2013.

[103] Liane E Hunter, Naomi Lubin, Nancy R Glassman, Xiaonan Xue, Moshe Spira, and Michael L Lipton. Comparing region of interest versus voxel-wise diffusion tensor imaging analytic methods in mild and moderate traumatic brain injury: A systematic review and meta-analysis. *Journal of neurotrauma*, 36(8):1222–1230, 2019.

[104] Khoi Minh Huynh, Geng Chen, Ye Wu, Dinggang Shen, and Pew-Thian Yap. Multi-site harmonization of diffusion mri data via method of moments. *IEEE transactions on medical imaging*, 38(7):1599–1609, 2019.

[105] Hyunho Hwang, Hafiz Zia Ur Rehman, and Sungon Lee. 3d u-net for skull stripping in brain mri. *Applied Sciences*, 9(3):569, 2019.

[106] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634, 2011.

[107] Matilde Inglese, Sachin Makani, Glyn Johnson, Benjamin A Cohen, Jonathan A Silver, Oded Gonen, and Robert I Grossman. Diffuse axonal injury in mild traumatic brain injury: a diffusion tensor imaging study. *Journal of neurosurgery*, 103(2):298–303, 2005.

[108] U Ito, H Tomita, Sh Yamazaki, Y Takada, and Y Inaba. Brain swelling and brain oedema in acute head injury. *Acta neurochirurgica*, 79(2-4):120–124, 1986.

[109] Nina Jacobsen, Andreas Deistung, Dagmar Timmann, Sophia L Goericke, Jürgen R Reichenbach, and Daniel Güllmar. Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network. *Zeitschrift für Medizinische Physik*, 29(2):128–138, 2019.

[110] Hamid A Jalab and A Hasan. Magnetic resonance imaging segmentation techniques of brain tumors: A review. *en), Arch Neurosci, Review Article vol. In Press, no. In Press, p. e84920*, 2019.

[111] Yasir N Jassam, Saef Izzy, Michael Whalen, Dorian B McGavern, and Joseph El Khoury. Neuroimmunology of traumatic brain injury: time for a paradigm shift. *Neuron*, 95(6):1246–1265, 2017.

[112] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, and Mark W Woolrich. Smith sm. *FSL neuroimage*, 62:782–90, 2012.

[113] Jens H Jensen and Joseph A Helpern. Mri quantification of non-gaussian water diffusion by kurtosis analysis. *NMR in Biomedicine*, 23(7):698–710, 2010.

[114] Ben Jeurissen, Jacques-Donald Tournier, Thijs Dhollander, Alan Connelly, and Jan Sijbers. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion mri data. *NeuroImage*, 103:411–426, 2014.

[115] John P John, Lei Wang, Amanda J Moffitt, Harmeeta K Singh, Mokhtar H Gado, and John G Csernansky. Inter-rater reliability of manual segmentation of the superior, inferior and middle frontal gyri. *Psychiatry Research: Neuroimaging*, 148(2-3):151–163, 2006.

[116] Victoria E Johnson, William Stewart, and Douglas H Smith. Axonal pathology in traumatic brain injury. *Experimental neurology*, 246:35–43, 2013.

[117] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[118] Derek K Jones and Peter J Basser. "squashing peanuts and smashing pumpkins": how noise distorts diffusion-weighted mr data. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 52(5):979–993, 2004.

[119] Jorge Jovicich, Frederik Barkhof, Claudio Babiloni, Karl Herholz, Christoph Mulert, Bart NM van Berckel, Giovanni B Frisoni, and SRA-NED JPND Working Group. Harmonization of neuroimaging biomarkers for neurodegenerative diseases: A survey in the imaging community of perceived barriers and suggested actions. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 11(C):69–73, 2019.

[120] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, et al. Mri-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects

of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage*, 46(1):177–192, 2009.

[121] Jorge Jovicich, Moira Marizzoni, Beatriz Bosch, David Bartrés-Faz, Jennifer Arnold, Jens Benninghoff, Jens Wiltfang, Luca Roccatagliata, Agnese Picco, Flavio Nobili, et al. Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. *Neuroimage*, 101:390–403, 2014.

[122] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

[123] Konstantinos Kamnitsas, Liang Chen, Christian Ledig, Daniel Rueckert, and Ben Glocker. Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. *Ischemic stroke lesion segmentation*, 13:46, 2015.

[124] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016.

[125] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[126] Suheyla Cetin Karayumak, Sylvain Bouix, Lipeng Ning, Anthony James, Tim Crow, Martha Shenton, Marek Kubicki, and Yogesh Rathi. Retrospective harmonization of multi-site diffusion mri data acquired with different acquisition parameters. *NeuroImage*, 184:180–200, 2019.

[127] Suheyla Cetin Karayumak, Marek Kubicki, and Yogesh Rathi. s. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 116–124. Springer, 2018.

[128] Terry Karpman, Sandra Wolfe, and James W Vargo. The psychological adjustment of adult clients and their parents following closed head injury. *J Rehabil Counsel*, 17:28–33, 1985.

[129] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003.

[130] Elias Kellner, Bibek Dhital, Valerij G Kiselev, and Marco Reisert. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magnetic resonance in medicine*, 76(5):1574–1581, 2016.

[131] NS King, S Crawford, FJ Wenden, NEG Moss, and DT Wade. The rivermead post concussion symptoms questionnaire: a measure of symptoms commonly experienced after head injury and its reliability. *Journal of neurology*, 242(9):587–592, 1995.

[132] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[133] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep mri brain extraction: a 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[134] Arno Klein, Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, D Louis Collins, James Gee, Pierre Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, 46(3):786–802, 2009.

[135] Simon Koppers, Luke Bloy, Jeffrey I Berman, Chantal MW Tax, J Christopher Edgar, and Dorit Merhof. Spherical harmonic residual network for diffusion signal harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–182. Springer, 2018.

[136] Simon Koppers and Dorit Merhof. Delimit pytorch-an extension for deep learning in diffusion imaging. *arXiv preprint arXiv:1808.01517*, 2018.

[137] Robert K Kosior, M Louis Lauzon, Nikolai Steffenhagen, Jayme C Kosior, Andrew Demchuk, and Richard Frayne. Atlas-based topographical scoring for magnetic resonance imaging of acute stroke. *Stroke*, 41(3):455–460, 2010.

[138] Marilyn F Kraus, Teresa Susmaras, Benjamin P Caughlin, Corey J Walker, John A Sweeney, and Deborah M Little. White matter integrity and cognition in chronic traumatic brain injury: a diffusion tensor imaging study. *Brain*, 130(10):2508–2519, 2007.

[139] Frithjof Kruggel, Jessica Turner, L Tugan Muftuler, Alzheimer's Disease Neuroimaging Initiative, et al. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the adni cohort. *Neuroimage*, 49(3):2123–2133, 2010.

[140] Dongyang Kuang and Tanya Schmah. Faim–a convnet method for unsupervised 3d medical image registration. *arXiv preprint arXiv:1811.09243*, 2018.

[141] Alok Kumar and David J Loane. Neuroinflammation after traumatic brain injury: opportunities for therapeutic intervention. *Brain, behavior, and immunity*, 26(8):1191–1201, 2012.

[142] Raj Kumar, Rakesh K Gupta, Mazhar Husain, Chaynika Chaudhry, Arti Srivastava, Sona Saksena, and Ram KS Rathore. Comparative evaluation of corpus callosum dti metrics in acute mild and moderate traumatic brain injury: its correlation with neuropsychometric tests. *Brain injury*, 23(7-8):675–685, 2009.

[143] David Kurland, Caron Hong, Bizhan Aarabi, Volodymyr Gerzanich, and J Marc Simard. Hemorrhagic progression of a contusion after traumatic brain injury: a review. *Journal of neurotrauma*, 29(1):19–31, 2012.

[144] Thomas Küstner, Karim Armanious, Jiahuan Yang, Bin Yang, Fritz Schick, and Sergios Gatidis. Retrospective correction of motion-affected mr images using deep learning frameworks. *Magnetic resonance in medicine*, 2019.

[145] Thomas Küstner, Annika Liebgott, Lukas Mauch, Petros Martirosian, Fabian Bamberg, Konstantin Nikolaou, Bin Yang, Fritz Schick, and Sergios Gatidis. Automated reference-free detection of motion artifacts in magnetic resonance images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 31(2):243–256, 2018.

[146] Emmanuel Lagarde, Louis-Rachid Salmi, Lena W Holm, Benjamin Contrand, Françoise Masson, Régis Ribéreau-Gayon, Magali Laborey, and J David Cassidy. Association of symptoms following mild traumatic brain injury with posttraumatic stress disorder vs postconcussion syndrome. *JAMA psychiatry*, 71(9):1032–1040, 2014.

[147] Bennett A Landman, Alan J Huang, Aliya Gifford, Deepti S Vikram, Issel Anne L Lim, Jonathan AD Farrell, John A Bogovic, Jun Hua, Min Chen, Samson Jarso, et al. Multi-parametric neuroimaging reproducibility: a 3-t resource study. *Neuroimage*, 54(4):2854–2866, 2011.

[148] Paul J Laurienti, Aaron S Field, Jonathan H Burdette, Joseph A Maldjian, Yi-Fen Yen, and Dixon M Moody. Dietary caffeine consumption modulates fmri measures. *Neuroimage*, 17(2):751–757, 2002.

[149] M Le, LYW Tang, E Hernández-Torres, M Jarrett, T Brosch, L Metz, DKB Li, A Traboulsee, RC Tam, A Rauscher, et al. Flair2 improves lesiontoads automatic segmentation of multiple sclerosis lesions in non-homogenized, multi-center, 2d clinical magnetic resonance images. *NeuroImage: Clinical*, page 101918, 2019.

[150] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[151] Christian Ledig, Rolf A Heckemann, Alexander Hammers, Juan Carlos Lopez, Virginia FJ Newcombe, Antonios Makropoulos, Jyrki Lötjönen, David K Menon, and Daniel Rueckert. Robust whole-brain segmentation: application to traumatic brain injury. *Medical image analysis*, 21(1):40–58, 2015.

[152] Daren Lee, Ivo Dinov, Bin Dong, Boris Gutman, Igor Yanovsky, and Arthur W Toga. Cuda optimization strategies for compute-and memory-bound neuroimaging algorithms. *Computer methods and programs in biomedicine*, 106(3):175–187, 2012.

[153] Hyeon Woo Lee, Mert R Sabuncu, and Adrian V Dalca. Few labeled atlases are necessary for deep-learning-based segmentation. *arXiv preprint arXiv:1908.04466*, 2019.

[154] Alia Lemkaddem, Alessandro Daducci, Serge Vulliemoz, Kieran O'Brien, François Lazeyras, Martinus Hauf, Roland Wiest, Reto Meuli, Margitta Seeck, Gunnar Krueger, et al. A multi-center study: intra-scan and inter-scan variability of diffusion spectrum imaging. *Neuroimage*, 62(1):87–94, 2012.

[155] Brian Levine, Sandra E Black, Gordon Cheung, Ann Campbell, Colleen O'Toole, and Michael L Schwartz. Gambling task performance in traumatic brain injury: relationships to injury severity, atrophy, lesion location, and cognitive and psychosocial outcome. *Cognitive and Behavioral Neurology*, 18(1):45–54, 2005.

[156] Josef M Ling, Amanda Pena, Ronald A Yeo, Flannery L Merideth, Stefan Klimaj, Charles Gasparovic, and Andrew R Mayer. Biomarkers of increased diffusion anisotropy in semi-acute mild traumatic brain injury: a longitudinal perspective. *Brain*, 135(4):1281–1292, 2012.

[157] Michael L Lipton, Edwin Gulko, Molly E Zimmerman, Benjamin W Friedman, Mimi Kim, Erik Gellella, Tamar Gold, Keivan Shifteh, Babak A Ardekani, and Craig A Branch. Diffusion-tensor imaging implicates prefrontal axonal injury in executive function impairment following very mild traumatic brain injury. *Radiology*, 252(3):816–824, 2009.

[158] Xianli Liu, Haifeng Zhao, Shaojie Zhang, and Zhenyu Tan. Brain image parcellation using multi-atlas guided adversarial fully convolutional network. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 723–726. IEEE, 2019.

[159] Xavier Lladó, Arnau Oliver, Mariano Cabezas, Jordi Freixenet, Joan C Vilanova, Ana Quiles, Laia Valls, Lluís Ramió-Torrentà, and Àlex Rovira. Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches. *Information Sciences*, 186(1):164–185, 2012.

[160] Benedikt Lorch, Ghislain Vaillant, Christian Baumgartner, Wenjia Bai, Daniel Rueckert, and Andreas Maier. Automated detection of motion artefacts in mr imaging using decision forests. *Journal of medical engineering*, 2017, 2017.

[161] Andrew IR Maas, David K Menon, P David Adelson, Nada Andelic, Michael J Bell, Antonio Belli, Peter Bragge, Alexandra Brazinova, András Büki, Randall M Chesnut, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *The Lancet Neurology*, 16(12):987–1048, 2017.

[162] Andrew IR Maas, David K Menon, Ewout W Steyerberg, Giuseppe Citerio, Fiona Lecky, Geoffrey T Manley, Sean Hill, Valerie Legrand, and Annina Sorgner. Collaborative european neurotrauma effectiveness research in traumatic brain injury (center-tbi) a prospective longitudinal observational study. *Neurosurgery*, 76(1):67–80, 2015.

[163] John D MacKenzie, Faez Siddiqi, James S Babb, Linda J Bagley, Lois J Mannon, Grant P Sinson, and Robert I Grossman. Brain atrophy in mild or moderate traumatic brain injury: a longitudinal quantitative analysis. *American Journal of Neuroradiology*, 23(9):1509–1515, 2002.

[164] Vincent A Magnotta, Joy T Matsui, Dawei Liu, Hans J Johnson, Jeffrey D Long, Bradley D Bolster Jr, Bryon A Mueller, Kelvin Lim, Susumu Mori, Karl G Helmer, et al. Multicenter reliability of diffusion tensor imaging. *Brain connectivity*, 2(6):345–355, 2012.

[165] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

[166] José V Manjón, Pierrick Coupé, Luis Concha, Antonio Buades, D Louis Collins, and Montserrat Robles. Diffusion weighted image denoising using overcomplete local pca. *PloS one*, 8(9):e73021, 2013.

[167] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[168] Stefano Marenco, Robert Rawlings, Gustavo K Rohde, Alan S Barnett, Robyn A Honea, Carlo Pierpaoli, and Daniel R Weinberger. Regional distribution of measurement error in diffusion tensor imaging. *Psychiatry Research: Neuroimaging*, 147(1):69–78, 2006.

[169] Ryan M Martin, Matthew J Wright, Evan S Lutkenhoff, Benjamin M Ellingson, John D Van Horn, Meral Tubi, Jeffry R Alger, David L McArthur, and Paul M Vespa. Traumatic hemorrhagic brain injury: impact of location and resorption on cognitive outcome. *Journal of neurosurgery*, 126(3):796–804, 2017.

[170] Makoto Matsushita, Kohkichi Hosoda, Yasuo Naitoh, Haruo Yamashita, and Eiji Kohmura. Utility of diffusion tensor imaging in the acute stage of mild to moderate traumatic brain injury for detecting white matter lesions and predicting long-term cognitive function in adults. *Journal of neurosurgery*, 115(1):130–139, 2011.

[171] AR Mayer, J Ling, MV Mannell, C Gasparovic, JP Phillips, D Doezema, R Reichard, and RA Yeo. A prospective diffusion tensor imaging study in mild traumatic brain injury. *Neurology*, 74(8):643–650, 2010.

[172] Ann C McKee and Daniel H Daneshvar. The neuropathology of traumatic brain injury. In *Handbook of clinical neurology*, volume 127, pages 45–66. Elsevier, 2015.

[173] DK Menon, K Schwab, DW Wright, and AI Maas. Demographics and clinical assessment working group of the international and interagency initiative toward common data elements for research on traumatic brain injury and psychological health. position statement: definition of traumatic brain injury. *Arch Phys Med Rehabil*, 91(11):1637–40, 2010.

[174] Arnaud Messé, Sophie Caplain, Gaëlle Paradot, Delphine Garrigue, Jean-François Mineo, Gustavo Soto Ares, Denis Ducreux, Frédéric Vignaud, Gaëlle Rozec, Hubert Desal, et al. Diffusion tensor imaging and white matter lesions at the subacute stage in mild traumatic brain injury with persistent neurobehavioral impairment. *Human brain mapping*, 32(6):999–1011, 2011.

[175] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523, 2016.

[176] Ludovico Minati and Władysław P Weglarz. Physical foundations, models, and methods of diffusion magnetic resonance imaging of the brain: A review. *Concepts in Magnetic Resonance Part A: An Educational Journal*, 30(5):278–307, 2007.

[177] Hengameh Mirzaalian, Amicie de Pierrefeu, Peter Savadjiev, Ofer Pasternak, Sylvain Bouix, Marek Kubicki, Carl-Fredrik Westin, Martha E Shenton, and Yogesh Rathi. Harmonizing diffusion mri data across multiple sites and scanners. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 12–19. Springer, 2015.

[178] Hengameh Mirzaalian, Lipeng Ning, Peter Savadjiev, Ofer Pasternak, Sylvain Bouix, O Michailovich, G Grant, CE Marx, Rajendra A Morey, LA Flashman, et al. Inter-site and inter-scanner diffusion mri data harmonization. *NeuroImage*, 135:311–323, 2016.

[179] Hengameh Mirzaalian, Lipeng Ning, Peter Savadjiev, Ofer Pasternak, Sylvain Bouix, Oleg Michailovich, Sarina Karmacharya, Gerald Grant, Christine E Marx, Rajendra A Morey, et al. Multi-site harmonization of diffusion mri data in a registration framework. *Brain imaging and behavior*, 12(1):284–295, 2018.

[180] Kent G Moen, Veronika Brezova, Toril Skandsen, Asta K Håberg, Mari Folvik, and Anne Vik. Traumatic axonal injury: the prognostic value of lesion load in corpus callosum, brain stem, and thalamus in different magnetic resonance imaging sequences. *Journal of neurotrauma*, 31(17):1486–1496, 2014.

[181] Susumu Mori, Setsu Wakana, Peter CM Van Zijl, and LM Nagae-Poetscher. *MRI atlas of human white matter*. Elsevier, 2005.

[182] Daniel Moyer, Greg Ver Steeg, Chantal MW Tax, and Paul M Thompson. Scanner invariant representations for diffusion mri harmonization. *Magnetic Resonance in Medicine*.

[183] Emma R Mulder, Remko A de Jong, Dirk L Knol, Ronald A van Schijndel, Keith S Cover, Pieter J Visser, Frederik Barkhof, Hugo Vrenken, Alzheimer's Disease Neuroimaging Initiative, et al. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods freesurfer and first. *Neuroimage*, 92:169–181, 2014.

[184] Ponnada A Narayana, Xintian Yu, Khader M Hasan, Elisabeth A Wilde, Harvey S Levin, Jill V Hunter, Emmy R Miller, Vipul Kumar S Patel, Claudia S Robertson, and James J McCarthy. Multi-modal mri of mild traumatic brain injury. *NeuroImage: Clinical*, 7:87–97, 2015.

[185] Abdullah Nazib, Clinton Fookes, and Dimitri Perrin. A comparative analysis of registration tools: Traditional vs deep learning approach on high resolution tissue cleared data. *arXiv preprint arXiv:1810.08315*, 2018.

[186] Virginia FJ Newcombe, Marta M Correia, Christian Ledig, Maria G Abate, Joanne G Outtrim, Doris Chatfield, Thomas Geeraerts, Anne E Manktelow, Eleftherios Garyfallidis, John D Pickard, et al. Dynamic changes in white matter abnormalities correlate with late improvement and deterioration following tbi: a diffusion tensor imaging study. *Neurorehabilitation and neural repair*, 30(1):49–62, 2016.

[187] Lipeng Ning, Elisenda Bonet-Carne, Francesco Grussu, Farshid Sepehrband, Enrico Kaden, Jelle Veraart, Stefano B Blumberg, Can Son Khoo, Marco Palombo, Iasonas Kokkinos, et al. Cross-scanner and cross-protocol multi-shell diffusion mri data harmonization: algorithms and results. *NeuroImage*, page 117128, 2020.

[188] Sumit N Niogi and Pratik Mukherjee. Diffusion tensor imaging of mild traumatic brain injury. *The Journal of head trauma rehabilitation*, 25(4):241–255, 2010.

[189] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[190] Ipek Oguz, Mahshid Farzinfar, Joy Matsui, Francois Budin, Zhexing Liu, Guido Gerig, Hans J Johnson, and Martin Andreas Styner. Dtiprep: quality control of diffusion-weighted images. *Frontiers in neuroinformatics*, 8:4, 2014.

[191] Stuart Oldham, Aurina Arnatkeviciute, Robert E Smith, Jeggan Tiego, Mark A Bellgrove, and Alex Fornito. The efficacy of different preprocessing steps in reducing motion-related confounds in diffusion mri connectomics. *bioRxiv*, 2020.

[192] William Ollier, Tim Sprosen, and Tim Peakman. Uk biobank: from concept to reality. 2005.

[193] Eva M Palacios, Alastair J Martin, Michael A Boss, Frank Ezekiel, Yi Shin Chang, Esther L Yuh, Mary J Vassar, David M Schnyer, Christine L MacDonald, Karen L Crawford, et al. Toward precision and reproducibility of diffusion tensor imaging: a multicenter diffusion phantom and traveling volunteer study. *American Journal of Neuroradiology*, 38(3):537–545, 2017.

[194] Nico Dario Papinutto, Francesca Maule, and Jorge Jovicich. Reproducibility and biases in high field brain diffusion mri: An evaluation of acquisition and analysis variables. *Magnetic resonance imaging*, 31(6):827–839, 2013.

[195] Kamlesh Pawar, Zhaolin Chen, N Jon Shah, and Gary F Egan. Motion correction in mri using deep convolutional neural network. In *Proceedings of the ISMRM Scientific Meeting & Exhibition, Paris*, volume 1174, 2018.

[196] Matthew L Pearn, Ingrid R Niesman, Junji Egawa, Atsushi Sawada, Angels Almenar-Queralt, Sameer B Shah, Josh L Duckworth, and Brian P Head. Pathophysiology associated with traumatic brain injury: current treatments and potential novel therapeutics. *Cellular and molecular neurobiology*, 37(4):571–585, 2017.

[197] Carlo Pierpaoli and Peter J Basser. Toward a quantitative assessment of diffusion anisotropy. *Magnetic resonance in Medicine*, 36(6):893–906, 1996.

[198] Maíra Siqueira Pinto, Roberto Paolella, Thibo Billiet, Pieter Van Dyck, Pieter-Jan Guns, Ben Jeurissen, Annemie Ribbens, Arnold J den Dekker, and Jan Sijbers. Harmonization of brain diffusion mri: Concepts and methods. *Frontiers in Neuroscience*, 14, 2020.

[199] Eric Plitman, Aurelie Bussy, Vanessa Valiquette, Alyssa Salaciak, Raihaan Patel, Marie-Lise Béland, Stephanie Tullo, Christine Tardif, M Natasha Rajah, Jamie Near, et al. The impact of the siemens trio to prisma upgrade and volumetric navigators on mri indices: A reliability study with implications for longitudinal study designs. *bioRxiv*, 2020.

[200] Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 17(11):1510, 2014.

[201] Jonathan R Polimeni and Kâmil Uludağ. Neuroimaging with ultra-high field mri: Present and future, 2018.

[202] Jennie L Ponsford, Marina G Downing, John Olver, Michael Ponsford, Rose Acher, Meagan Carty, and Gershon Spitz. Longitudinal follow-up of patients with traumatic brain injury: outcome at two, five, and ten years post-injury. *Journal of Neurotrauma*, 31(1):64–77, 2014.

[203] V Pop and J Badaut. A neurovascular perspective for long-term changes after brain trauma. *Translational stroke research*, 2(4):533–545, 2011.

[204] Matthew R Powell, Allen W Brown, Danielle Klunk, Jennifer R Geske, Kamini Krishnan, Cassie Green, and Thomas F Bergquist. Injury severity and depressive symptoms in a post-acute brain injury rehabilitation sample. *Journal of clinical psychology in medical settings*, 26(4):470–482, 2019.

[205] Ferran Prados, M Jorge Cardoso, Niamh Cawley, Baris Kanber, Olga Ciccarelli, Claudia AM Gandini Wheeler-Kingshott, and Sébastien Ourselin. Fully automated patch-based image restoration: Application to pathology inpainting. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 3–15. Springer, 2016.

[206] Charlotte L Rae, Marta M Correia, Ellemarije Altena, Laura E Hughes, Roger A Barker, and James B Rowe. White matter pathology in parkinson's disease: the effect of imaging protocol differences and relevance to executive function. *Neuroimage*, 62(3):1675–1684, 2012.

[207] Martin Rajchl, Nick Pawlowski, Daniel Rueckert, Paul M Matthews, and Ben Glocker. Neuronet: Fast and robust reproduction of multiple brain image segmentation pipelines. *arXiv preprint arXiv:1806.04224*, 2018.

[208] Jaime Ramos-Cejudo, Thomas Wisniewski, Charles Marmar, Henrik Zetterberg, Kaj Blennow, Mony J de Leon, and Silvia Fossati. Traumatic brain injury and alzheimer's disease: the cerebrovascular link. *EBioMedicine*, 28:21–30, 2018.

[209] Gabriel Ramos-Llordén, Gonzalo Vegas-Sánchez-Ferrero, Congyu Liao, Carl-Fredrik Westin, Kawin Setsompop, and Yogesh Rathi. Snr-enhanced diffusion mri with structure-preserving low-rank denoising in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2009.06600*, 2020.

[210] Marin E Ranta, Deana Crocetti, Jacqueline A Clauss, Michael A Kraut, Stewart H Mostofsky, and Walter E Kaufmann. Manual mri parcellation of the frontal lobe. *Psychiatry Research: Neuroimaging*, 172(2):147–154, 2009.

[211] Logan R Ranzenberger and Travis Snyder. Diffusion tensor imaging. In *StatPearls [Internet]*. StatPearls Publishing, 2019.

[212] Amna Rehman and Yasir Al Khalili. Neuroanatomy, occipital lobe. 2019.

[213] Lal Rehman, Ali Afzal, Hafiza Fatima Aziz, Sana Akbar, Asad Abbas, and Raza Rizvi. Radiological parameters to predict hemorrhagic progression of traumatic contusional brain injury. *Journal of neurosciences in rural practice*, 10(2):212, 2019.

[214] Florence CM Reith, Anneliese Synnot, Ruben van den Brande, Russell L Gruen, and Andrew IR Maas. Factors influencing the reliability of the glasgow coma scale: a systematic review. *Neurosurgery*, 80(6):829–839, 2017.

[215] Sophie Richter, Stefan Winzeck, Evgenios N Kornaropoulos, Tilak Das, Thijs Vande Vyvere, Jan Verheyden, Guy B Williams, Marta M Correia, David K Menon, Virginia FJ Newcombe, et al. Neuroanatomical substrates and symptoms associated with magnetic resonance imaging of patients with mild traumatic brain injury. *JAMA Network Open*, 4(3):e210994–e210994, 2021.

[216] H Robert, B Pichon, and R Haddad. Sexual dysfunctions after traumatic brain injury: Systematic review of the literature. *Progres en urologie: journal de l'Association francaise d'urologie et de la Societe francaise d'urologie*, 29(11):529–543, 2019.

[217] Bob Roozenbeek, Andrew IR Maas, and David K Menon. Changing patterns in the epidemiology of traumatic brain injury. *Nature Reviews Neurology*, 9(4):231, 2013.

[218] David E Ross, Alfred L Ochs, Jan M Seabaugh, Michael F DeMark, Carole R Shrader, Jennifer H Marwitz, and Michael D Havranek. Progressive brain atrophy in patients with chronic neuropsychiatric symptoms after mild traumatic brain injury: a preliminary study. *Brain injury*, 26(12):1500–1509, 2012.

[219] David E Ross, John D Seabaugh, Jan M Seabaugh, Claudia Alvarez, Laura Peyton Ellis, Christopher Powell, Christopher Hall, Christopher Reese, Leah Cooper, and Alfred L Ochs. Patients with chronic mild or moderate traumatic brain injury have abnormal brain enlargement. *Brain injury*, 34(1):11–19, 2020.

[220] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186:713–727, 2019.

[221] Snehashis Roy, Andrew Knutsen, Alexandru Korotcov, Asamoah Bosomtwi, Bernard Dardzinski, John A Butman, and Dzung L Pham. A deep learning framework for brain extraction in humans and animals with traumatic brain injury. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 687–691. IEEE, 2018.

[222] Gholamreza Salimi-Khorshidi, Stephen M Smith, John R Keltner, Tor D Wager, and Thomas E Nichols. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage*, 45(3):810–823, 2009.

[223] Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, James Loughead, Karthik Prabhakaran, Monica E Calkins, Ryan Hopson, Chad Jackson, Jack Keefe, Marisa Riley, et al. Neuroimaging of the philadelphia neurodevelopmental cohort. *Neuroimage*, 86:544–553, 2014.

[224] David J Schretlen and Anne M Shapiro. A quantitative review of the effects of traumatic brain injury on cognitive functioning. *International review of psychiatry*, 15(4):341–349, 2003.

[225] Johanna Seitz, Suheyla Cetin-Karayumak, Amanda Lyall, Ofer Pasternak, Madhura Baxi, Mark Vangel, Godfrey Pearlson, Carol Tamminga, John Sweeney, Brett Clementz, et al.

Investigating sexual dimorphism of human white matter in a harmonized, multisite diffusion magnetic resonance imaging study. *Cerebral cortex*.

[226] Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, et al. The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1):204, 2014.

[227] David J Sharp and Timothy E Ham. Investigating white matter injury after mild traumatic brain injury. *Current opinion in neurology*, 24(6):558–563, 2011.

[228] Teena Shetty, Joseph T Nguyen, Taylor Cogsil, Apostolos John Tsiouris, Sumit N Niogi, Esther U Kim, Aashka Dalal, Kristin Halvorsen, Kelianne Cummings, Tianhao Zhang, et al. Clinical findings in a multicenter mri study of mild tbi. *Frontiers in neurology*, 9:836, 2018.

[229] Dennis W Simon, Mandy J McGeachy, Hülya Bayır, Robert SB Clark, David J Loane, and Patrick M Kochanek. The far-reaching scope of neuroinflammation after traumatic brain injury. *Nature Reviews Neurology*, 13(3):171, 2017.

[230] Thamil Mani Sivanandam and Mahendra Kumar Thakur. Traumatic brain injury: a risk factor for alzheimer's disease. *Neuroscience & Biobehavioral Reviews*, 36(5):1376–1381, 2012.

[231] Stephen M Smith. Fast robust automated brain extraction.

[232] Stephen M Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E Nichols, Clare E Mackay, Kate E Watkins, Olga Ciccarelli, M Zaheer Cader, Paul M Matthews, et al. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31(4):1487–1505, 2006.

[233] Stephen M Smith and Thomas E Nichols. Statistical challenges in "big data" human neuroimaging. *Neuron*, 97(2):263–268, 2018.

[234] Samuel St-Jean, Max A Viergever, and Alexander Leemans. Harmonization of diffusion mri datasets with adaptive dictionary learning. *arXiv preprint arXiv:1910.00272*, 2019.

[235] Sheeba J Sujit, Refaat E Gabr, Ivan Coronado, Melvin Robinson, Sushmita Datta, and Ponnada A Narayana. Automated image quality evaluation of structural brain magnetic resonance images using deep convolutional neural networks. In *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*, pages 33–36. IEEE, 2018.

[236] Jean Talairach. Co-planar stereotaxic atlas of the human brain-3-dimensional proportional system. *An approach to cerebral imaging*, 1988.

[237] Ryutaro Tanno, Daniel E Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer with cnns:

exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.

[238] Chantal MW Tax, Francesco Grussu, Enrico Kaden, Lipeng Ning, Umesh Rudrapatna, C John Evans, Samuel St-Jean, Alexander Leemans, Simon Koppers, Dorit Merhof, et al. Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms. *NeuroImage*, 195:285–299, 2019.

[239] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269, 2017.

[240] Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304(7872):81–84, 1974.

[241] Stefan J Teipel, Sigrid Reuter, Bram Stieltjes, Julio Acosta-Cabronero, Ulrike Ernemann, Andreas Fellgiebel, Massimo Filippi, Giovanni Frisoni, Frank Hentschel, Frank Jessen, et al. Multicenter stability of diffusion tensor imaging measures: a european clinical and physical phantom study. *Psychiatry Research: Neuroimaging*, 194(3):363–371, 2011.

[242] Paul Thompson. Enigma, big data, and neuroimaging genetics in 50,000 people from 35 countries: Challenges and lessons learned. *European Neuropsychopharmacology*, 29:S769–S770, 2019.

[243] Neil J Tolentino, Christina E Wierenga, Shana Hall, Susan F Tapert, Martin P Paulus, Thomas T Liu, Tom L Smith, and Marc A Schuckit. Alcohol effects on cerebral blood flow in subjects with low and high responses to alcohol. *Alcoholism: Clinical and Experimental Research*, 35(6):1034–1040, 2011.

[244] Qiqi Tong, Ting Gong, Hongjian He, Zheng Wang, Wenwen Yu, Jianjun Zhang, Lihao Zhai, Hongsheng Cui, Xin Meng, Chantal WM Tax, et al. A deep learning–based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols. *Magnetic Resonance Imaging*, 73:31–44, 2020.

[245] Qiqi Tong, Hongjian He, Ting Gong, Chen Li, Peipeng Liang, Tianyi Qian, Yi Sun, Qiuping Ding, Kuncheng Li, and Jianhui Zhong. Reproducibility of multi-shell diffusion tractography on traveling subjects: A multicenter study prospective. *Magnetic resonance imaging*, 59:1–9, 2019.

[246] Arnold Toth. Magnetic resonance imaging application in the area of mild and acute traumatic brain injury. In *Brain neurotrauma: molecular, neuropsychological, and rehabilitation aspects*. CRC Press/Taylor & Francis, 2015.

[247] Arnold Toth, Noemi Kovacs, Gabor Perlaki, Gergely Orsi, Mihaly Aradi, Hedvig Komaromy, Erzsebet Ezer, Peter Bukovics, Orsolya Farkas, Jozsef Janszky, et al. Multi-modal magnetic resonance imaging in the acute and sub-acute phase of mild traumatic brain injury: can we see the difference? *Journal of neurotrauma*, 30(1):2–10, 2013.

[248] Jeffrey Tsao and Sebastian Kozerke. Mri temporal acceleration techniques. *Journal of Magnetic Resonance Imaging*, 36(3):543–560, 2012.

[249] Jessica A Turner. The rise of large-scale imaging studies in psychiatry. *GigaScience*, 3(1):2047–217X, 2014.

[250] Marleen Maria van Eijck, Guus Geurt Schoonman, Joukje van der Naalt, Jolanda de Vries, and Gerwin Roks. Diffuse axonal injury after traumatic brain injury is a prognostic factor for functional outcome: a systematic review and meta-analysis. *Brain injury*, 32(4):395–402, 2018.

[251] David C Van Essen, Kamil Ugurbil, E Auerbach, D Barch, TEJ Behrens, R Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[252] John Darrell Van Horn and Arthur W Toga. Human neuroimaging as a "big data" science. *Brain imaging and behavior*, 8(2):323–331, 2014.

[253] Vigneswaran Veeramuthu, Vairavan Narayanan, Tan Li Kuo, Lisa Delano-Wood, Karuthan Chinna, Mark William Bondi, Vicknes Waran, Dharmendra Ganesan, and Norlisah Ramli. Diffusion tensor imaging parameters in mild traumatic brain injury and its correlation with early neuropsychological impairment: a longitudinal study. *Journal of neurotrauma*, 32(19):1497–1509, 2015.

[254] Jelle Veraart, Els Fieremans, Ileana O Jelescu, Florian Knoll, and Dmitry S Novikov. Gibbs ringing in diffusion mri. *Magnetic resonance in medicine*, 76(1):301–314, 2016.

[255] Jelle Veraart, Dmitry S Novikov, Daan Christiaens, Benjamin Ades-Aron, Jan Sijbers, and Els Fieremans. Denoising of diffusion mri using random matrix theory. *NeuroImage*, 142:394–406, 2016.

[256] Christian Vollmar, Jonathan O'Muircheartaigh, Gareth J Barker, Mark R Symms, Pamela Thompson, Veena Kumari, John S Duncan, Mark P Richardson, and Matthias J Koepp. Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 t scanners. *Neuroimage*, 51(4):1384–1394, 2010.

[257] Bodil C Vos, Karen Nieuwenhuijsen, and Judith K Sluiter. Consequences of traumatic brain injury in professional american football players: a systematic review of the literature. *Clinical journal of sport medicine*, 28(2):91–99, 2018.

[258] Christian Wachinger, Anna Rieckmann, and Sebastian Pölsterl. Detect and correct bias in multi-site neuroimaging datasets. *arXiv preprint arXiv:2002.05049*, 2020.

[259] Erica J Wallace, Jane L Mathias, and Lynn Ward. Diffusion tensor imaging changes following mild, moderate and severe adult traumatic brain injury: a meta-analysis. *Brain imaging and behavior*, 12(6):1607–1621, 2018.

[260] Yi Wang, Aditya Gupta, Zhexing Liu, Hui Zhang, Maria L Escolar, John H Gilmore, Sylvain Gouttard, Pierre Fillard, Eric Maltbie, Guido Gerig, et al. Dti registration in atlas based fiber analysis of infantile krabbe disease. *Neuroimage*, 55(4):1577–1586, 2011.

[261] Jeffrey B Ware, Tessa Hart, John Whyte, Amanda Rabinowitz, John A Detre, and Junghoon Kim. Inter-subject variability of axonal injury in diffuse traumatic brain injury. *Journal of neurotrauma*, 34(14):2243–2253, 2017.

[262] Theodore Wasserman and Angela Mion. The progression of memory loss secondary to tbi-induced white matter attenuation: a review of the literature and case exemplar. *Journal of pediatric neuropsychology*, 5(1-2):31–40, 2019.

[263] Jakob Wasserthal, Peter Neher, and Klaus H Maier-Hein. Tractseg-fast and accurate white matter tract segmentation. *NeuroImage*, 183:239–253, 2018.

[264] Yenny Webb-Vargas, Shaojie Chen, Aaron Fisher, Amanda Mejia, Yuting Xu, Ciprian Crainiceanu, Brian Caffo, and Martin A Lindquist. Big data and neuroimaging. *Statistics in biosciences*, 9(2):543–558, 2017.

[265] Emerson M Wickwire, David M Schnyer, Anne Germain, Michael T Smith, Scott G Williams, Christopher J Lettieri, Ashlee B McKeon, Steven M Scharf, Ryan Stocker, Jennifer Albrecht, et al. " sleep, sleep disorders, and circadian health following mild traumatic brain injury in adults: Review and research agenda": Correction. 2019.

[266] V Wiggermann, E Hernandez-Torres, A Traboulsee, DKB Li, and A Rauscher. Flair2: a combination of flair and t2 for improved ms lesion detection. *American Journal of Neuroradiology*, 37(2):259–265, 2016.

[267] EA Wilde, SR McCauley, JV Hunter, ED Bigler, Z Chu, ZJ Wang, GR Hanten, M Troyanskaya, R Yallampalli, X Li, et al. Diffusion tensor imaging of acute mild traumatic brain injury in adolescents. *Neurology*, 70(12):948–955, 2008.

[268] Elisabeth A Wilde, Stephen R McCauley, Amanda Barnes, Trevor C Wu, Zili Chu, Jill V Hunter, and Erin D Bigler. Serial measurement of memory and diffusion tensor imaging changes within the first week following uncomplicated mild traumatic brain injury. *Brain imaging and behavior*, 6(2):319–328, 2012.

[269] JT Lindsay Wilson, Laura EL Pettigrew, and Graham M Teasdale. Structured interviews for the glasgow outcome scale and the extended glasgow outcome scale: guidelines for their use. *Journal of neurotrauma*, 15(8):573–585, 1998.

[270] Trevor C Wu, Elisabeth A Wilde, Erin D Bigler, Ragini Yallampalli, Stephen R McCauley, Maya Troyanskaya, Zili Chu, Xiaoqi Li, Gerri Hanten, Jill V Hunter, et al. Evaluating the relationship between memory functioning and cingulum bundles in acute mild traumatic brain injury using diffusion tensor imaging. *Journal of neurotrauma*, 27(2):303–307, 2010.

[271] Xin Wu, Ivan I Kirov, Oded Gonen, Yulin Ge, Robert I Grossman, and Yvonne W Lui. Mr imaging applications in mild traumatic brain injury: an imaging update. *Radiology*, 279(3):693–707, 2016.

[272] Andy Wai Kan Yeung. Most common publication types of neuroimaging literature: Papers with high levels of evidence are on the rise. *Frontiers in Human Neuroscience*, 14:136, 2020.

[273] Bo Yin, Dan-Dong Li, Huan Huang, Cheng-Hui Gu, Guang Hui Bai, Liu-Xun Hu, Jin-Fei Zhuang, and Ming Zhang. Longitudinal changes in diffusion tensor imaging following mild traumatic brain injury and correlation with outcome. *Frontiers in neural circuits*, 13:28, 2019.

[274] John K Yue, Mary J Vassar, Hester F Lingsma, Shelly R Cooper, David O Okonkwo, Alex B Valadka, Wayne A Gordon, Andrew IR Maas, Pratik Mukherjee, Esther L Yuh, et al. Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury. *Journal of neurotrauma*, 30(22):1831–1844, 2013.

[275] John K Yue, Ethan A Winkler, Ross C Puffer, Hansen Deng, Ryan RL Phelps, Sagar Wagle, Molly Rose Morrissey, Ernesto J Rivera, Sarah J Runyon, Mary J Vassar, et al. Temporal lobe contusions on computed tomography are associated with impaired 6-month functional recovery after mild traumatic brain injury: a track-tbi study. *Neurological research*, 40(11):972–981, 2018.

[276] Esther L Yuh, Shelly R Cooper, Pratik Mukherjee, John K Yue, Hester F Lingsma, Wayne A Gordon, Alex B Valadka, David O Okonkwo, David M Schnyer, Mary J Vassar, et al. Diffusion tensor imaging for outcome prediction in mild traumatic brain injury: a track-tbi study. *Journal of neurotrauma*, 31(17):1457–1477, 2014.

[277] Evangelia I Zacharaki, Stathis Kanterakis, R Nick Bryan, and Christos Davatzikos. Measuring brain lesion progression with a supervised tissue classification system. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 620–627. Springer, 2008.

[278] Lyubomir Zagorchev, Carsten Meyer, Thomas Stehle, Fabian Wenzel, Stewart Young, Jochen Peters, Juergen Weese, Keith Paulsen, Matthew Garlinghouse, James Ford, et al. Differences

in regional brain volumes two months and one year after mild traumatic brain injury. *Journal of neurotrauma*, 33(1):29–34, 2016.

[279] Liang Zhan, Alex D Leow, Neda Jahanshad, Ming-Chang Chiang, Marina Barysheva, Agatha D Lee, Arthur W Toga, Katie L McMahon, Greig I De Zubicaray, Margaret J Wright, et al. How does angular resolution affect diffusion imaging measures? *Neuroimage*, 49(2):1357–1371, 2010.

[280] Hui Zhang, Brian B Avants, Paul A Yushkevich, John H Woo, Sumei Wang, Leo F McCluskey, Lauren B Elman, Elias R Melhem, and James C Gee. High-dimensional spatial normalization of diffusion tensor images improves the detection of white matter differences: an example study using amyotrophic lateral sclerosis. *IEEE transactions on medical imaging*, 26(11):1585–1597, 2007.

[281] Hui Zhang, Paul A Yushkevich, Daniel C Alexander, and James C Gee. Deformable registration of diffusion tensor mr images with explicit orientation optimization. *Medical image analysis*, 10(5):764–785, 2006.

[282] Hui Zhang, Paul A Yushkevich, Daniel Rueckert, and James C Gee. Unbiased white matter atlas construction using diffusion tensor images. In *International conference on medical image computing and computer-assisted intervention*, pages 211–218. Springer, 2007.

[283] Xiaopeng Zhou, Ken E Sakaie, Josef P Debbins, Sridar Narayanan, Robert J Fox, and Mark J Lowe. Scan-rescan repeatability and cross-scanner comparability of dti metrics in healthy subjects in the sprint-ms multicenter trial. *Magnetic resonance imaging*, 53:105–111, 2018.

[284] Yongxia Zhou, Andrea Kierans, Damon Kenul, Yulin Ge, Joseph Rath, Joseph Reaume, Robert I Grossman, and Yvonne W Lui. Mild traumatic brain injury: longitudinal regional brain volume changes. *Radiology*, 267(3):880–890, 2013.

[285] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.